智慧医疗的机器学习模型可解释性方法研究综述

摘要：

深度学习模型强大的表现力，将我们带入一个“智能”时代，同时也因为它的不透明性和不可解释性，停下了快速发展的步伐。可解释性是打开深度学习这个黑盒子的钥匙，各领域专家对可解释性有不同的见解，也提出了大量可解释性方法，不同的评定标准。各种可解释性方法有其特点和优缺点，在现有研究上并提出优化的方法。可解释性方法在图像处理和自然语言处理领域取得了一些成果，但是在一些需要高解释性和高安全性的领域，例如自动驾驶、智慧医疗和网络安全等，需要考虑这些领域独有的特征，提出切合的可解释性方法。在本文中，阐述可解释性对于机器学习的重要性，以及现有研究的可解释性方法，并给出比较和讨论。以及对可解释性问题的相关启发。

关键字：机器学习模型，深度学习，可解释性，解释方法，智慧医疗。

1.引言

从 1956 年在美国达特茅斯学院开会研讨“如何用机器模拟人的智能”，首次提出“人工智能（Artificial Intelligence，简称 AI）”这一概念，标志着人工智能学科的诞生。在此次会议之后，人工智能迎来了第一波快速发展时期，也让专家们看到了机器走向智能的希望。随着社会的进步，人工智能面临了三大问题：计算能力不足、问题复杂性增加、数据量的缺失，让人工智能放缓了发展的脚步。1980 年卡内基梅隆大学为数字设备公司设计了一套名为 XCON 的“专家系统”，实现了人工智能从理论研究走向实际应用、从一般推理策略探讨转向运用专门知识的重大突破，但是上述的三个问题仍然没有得到解决。上世界九十年代，随着 AI 技术的发展，尤其是神经网络的发展，人们开始对人工智能有客观理性的认识，人工智能的研究领域主要分为五个方面：最底层的基础设施（硬件/算力、大数据），在此基础上是算法（机器学习、深度学习），技术方向，具体技术以及行业解决方案。人工智能作为新一轮科技变革的核心力量，推动传统产

业的升级换代，逐步转化为“智能时代”。但同时也会带来很多不可避免的问题，人工智能的透明性和不可解释性也逐步凸显出来。

机器学习是人工智能一个重要分支，是当前解决大多数人工智能问题的核心基石，机器学习的核心思想是使计算机模拟人学习的过程，它将从样本数据中学习到知识和规律，然后学得的模型用于实际的推断和决策，它与普通程序的区别在于这是一种数据驱动的方法。机器学习分为三类，第一类无监督学习，让机器自主从数据中学习规律，并分为各个类别。第二类是监督学习，给定数据样本和标签，经过模型的不断训练，得到模型进而预测结果。第三类是强化学习，在没有任何标签的情况下，通过先尝试做出一些行为，然后得到结果，这个结果就是给机器一个反馈，然后通过这个反馈再来调整。强化学习和监督学习都是从样本数据到输出的一个映射，监督学习得到的是他们之间的关系，而强化学习得到的是机器的反馈。而非监督式学习从数据中学习到的是一种模式。传统的机器学习算法包括贝叶斯，决策树，推理逻辑等，需要人为的找寻数据中的特征，也就是特征的准确与否决定了学习效果的好坏，而特征选择需要大量的先验知识来确定，也就不能实现完全意义上的人工智能。

深度学习是一种特殊的机器学习，它的引入是为了更好的实现人工智能。它是由多层神经网络构成。神经网络由输入层、隐藏层（单层或多层）和输出层构成。而隐藏层由多个神经元组成，神经网络的训练实际是神经元与神经元之间参数的调整，即权重。相比传统的机器学习，神经网络的激活函数是会进行非线性转换，来提高网络的表达能力，处理线性不可分的问题。深度学习在语音识别，图像分类等问题中都有很强的表现力，它本质上是在拟合函数，而这个函数通常是非线性的，这就意味着不不能直观的理解深度神经网络做出的决定，不能了解它的整个决策过程。所以在我们看来深度神经网路就像一个“黑盒子”，我们只知道它性能良好，而不知它到底是怎么好，是如何做出这种决策，随之而来的问题是我们能否相信机器做出的决定？机器做出的决定是否正确？正因为这种不可解释性，深度学习在很多领域的应用不能落实，也无法推动人工智能再向前发展。那么对机器学习模型做出解释成为当下最重要的问题之一。

医疗问题是与我们生活息息相关的，如何提高就诊效率和准确度，给出最有效的治疗方案，为病人争取宝贵的时间，这是医生和病人最关心的问题。根据美

国心脏协会（AHA）对心脏病和卒中流行病学统计数据 2019 版报告显示，心血管疾病是全球主要的死亡原因，2016 年全球死亡人数超过 1760 万，预计到 2030 年将会增加到 2360 万以上。根据中国心血管病中心发布的《中国心血管病报告 2018》[41]，中国心血管病（CVD）患病率处于持续上升阶段，推算 CVD 现患人数 2.9 亿，其死亡率居于首位，高于肿瘤及其他疾病。医生人数与病患人数的巨大差异，导致一大部分患者得不到及时的救治，错失治疗时间。将机器学习的方法应用于心电图诊断，将会在很大程度上缓解上述问题，但同时也因为机器学习的不可解释性和不透明性，难以大规模应用于临床实践。

基于上述面临的挑战和问题，本文将讲述机器学习可解释性的相关问题以及在心电领域的应用，并比较现存多数方法与心电应用的不适应与差别，进行归纳总结。

2.可解释性

什么是机器学习的可解释性？可解释性是一种以人能够理解的方式表达机器的学习过程，是人与机器建立信任的一个桥梁。自动驾驶从提出到现在任然没有实现完全的自主驾驶，2017 年 Uber 在进行道路测试时发生交通事故致人死亡，自动驾驶使用深度学习模型，使用大量交通数据经过长时间训练，从而完成决策任务，这其中用到的深度神经网络，规模之大，甚至是数以万计的参数，我们不能准确知道模型是依据什么来做出最后的决策。这种情况下需要可解释性的。广告预测，商品推荐等应用场景，对可解释性的要求并不高。即使模型出现预测错误，并不会产生对人身巨大伤害的后果。如果不能做到完全信任深度学习，也就无法发挥深度学习的长处，解决实际问题。考虑模型的可解释性的同时，模型的准确度也需要权衡。模型越复杂，那么它的学习能力越强大，会更好的学习数据样本中的知识和规律，同时它的可解释性就越低，反之，模型简单，准确率低，可解释性强。由此，我们针对不同的实际问题，要在这两者中做出取舍，是选择简单可解释性强的模型完成训练任务，还是复杂可解释性差的模型然后针对模型提出可解释方法？可以将可解释性方法大致分为两类，一类是针对模型本身进行解释，第二类是借助解释模块或者解释模型进行解释。[1]

2.1 模型本身解释方法

模型本身解释方法是指通过模型本身的信息来进行解释，分为三种不同的形

式。

2.1.1 数据探索

机器学习是从大量数据中学习到数据的相互联系和影响，如图 1 所示，通过数据预处理和数据可视化[6]的方法，从而理解神经网络在每一层学到的特征。在选择合适的模型之前对数据的大致了解，知道数据的分布情况或者对特征进行探索[5]，帮助我们找到代表性的或者不具代表性的样本[8]（如图 2 所示），通过 Influence Function 来判断分析哪些样本是有利于模型的，哪些样本是对模型不起决定性作用，进而提高模型的准确性和可靠性。当数据量大或者数据维度非常高时，而数据可视化能够给我们一个关于数据更加直观的感受，帮助从多个维度去了解数据的特征，有助于我们找到更加合适的模型。

通过上述方法，实验证明在一定程度上提高了模型的准确度并给出了人能够直观理解的解释，Koh P W 等人[8]提出的方法是针对数据样本点，是一种局部的解释方法。基于数据探索的解释方法，是贴合深度学习模型的，不会损害模型准确度，并能够通过这样的方法对数据样本进行一个初步筛选，从而达到提升模型的目的。

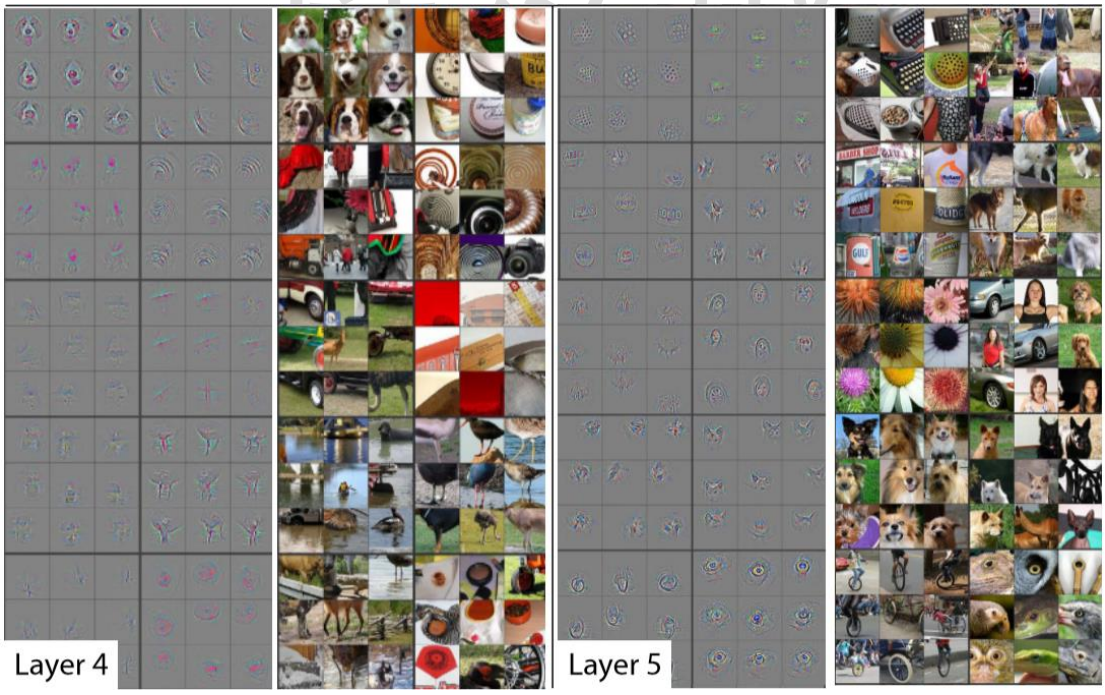

图 1[6]神经网络中的第 4 层和第 5 层，作为可视化的表达。随机选取验证集中特征图的前 9 个最大激活，并使用反卷积网络的方法将其投影到像素空间。

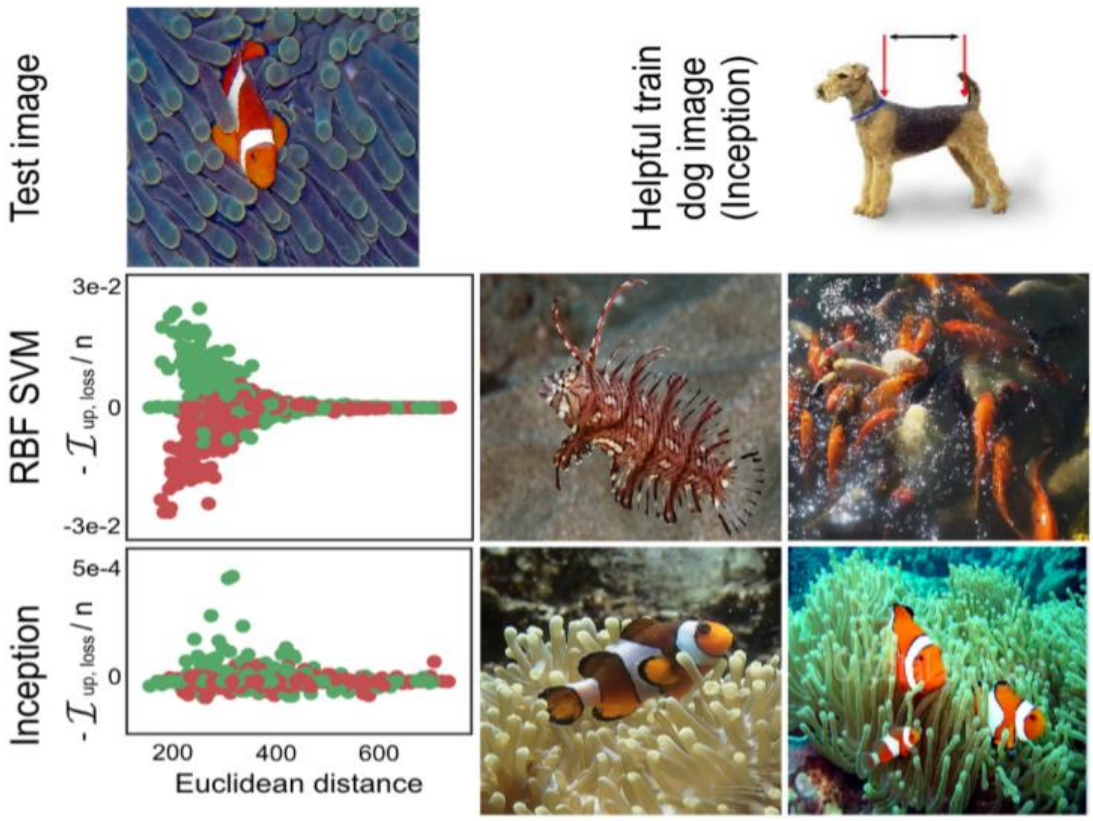

图 2 左图红色的点代表狗，绿色的点代表鱼。右图的上面是训练集中一张狗的图片，在 Inception 模型中被认为是帮助模型正确分类为鱼，右下图是对两个模型都最有帮助的图片。

2.1.2 可解释性的模型

线性模型，在朴素贝叶斯模型[3]中，是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，学习从输入到输出的联合概率分布，可以通过特征的概率来解释预测结果。在线性模型中，利用数理统计中分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，通过分析模型的参数来反映样本特征的重要性，如果样本权重的绝对值越大，那么这个样本对模型的贡献越大，反之则越小。在决策树模型中，从根节点开始，每一个节点都代表一个属性，最后的叶子节点即为分类结果，而且它是一种简单非参数的模型，不需要对数据有任何的先验假设，计算速度快，且易于解释，但是构造有许多分支的决策树，是复杂且耗时的，并且太多的分支线索会导致可解释下降。基于规则的模型，是用 if-else 来完成一步步的规则分解。

上述模型的复杂度偏低，相比深度学习模型，他们在复杂任务的情况下，预测能力弱，因此在线性模型的基础上，有了广义加性模型[1]，保留线性模型本

身的高解释性，同时也提高了模型的表达能力。Wu M[11]等提出的方法，在决策树的基础上增加正则化，减少决策树的节点，在不损失模型准确度的情况下解决了决策树节点庞大的问题。这类模型最重要的是根据预测任务平衡准确度和模型可解释性，适合简单的特征相关性不高的预测任务，达到双赢的目的。

2.1.3 解释模块

是通过引入额外的解释模块来实现，注意力机制是其中的一种方法，注意力机制从本质上讲和人类的选择性视觉注意力机制类似，核心目标也是从众多信息中选择性专注于出对当前任务目标更关键的信息，忽视其他不相关的信息。对通过分析注意力权重来对结果做出解释，从而增加了模型的可解释性。注意力机制在图像识别、自然语言处理以及推荐系统中得到了广泛的应用[32]，也逐渐在医疗领域得到应用。Choi E[33]等人提出的 RNN 结合两层注意力机制用于顺序数据，该方法对预测结果给出了详细的解释，并且保留了 RNN 相对的精度，在学习可解释表示的同时，使用注意力机制生成时间序列信息，模仿医生的就诊行为，以相反的时间顺序检查患者过去的就诊记录，从而促进更稳定的注意力机制的完善。

另一种方法是构建知识图谱，知识图谱具有海量规模，语义丰富，结构友好的特点，机器可以通过理解关键字，从而实现从搜索直接通往答案，做到精确的分析，并给出很好的解释性。根据知识图谱的类别可以大致分为普通知识图谱和专业知识图谱，普通知识图谱是与我们日常生活相关的各个方面，专业性弱构造难度小，专业知识图谱需要大量的专业知识，专业性强构造难度大。因此对于知识图谱的构建任然存在很大的困难，如何将知识转化为符号融入进机器学习中，如何让机器学会所谓的常识推理等问题。知识图谱多应用于推荐系统，搜索等[36]，鲜少应用于医疗领域。

引入额外的解释模块相比可视化数据以及特征探索，更适合复杂的网络结构，与可解释模型相比，这类方法在不损失模型准确度的情况下提供了很好的可解释性。

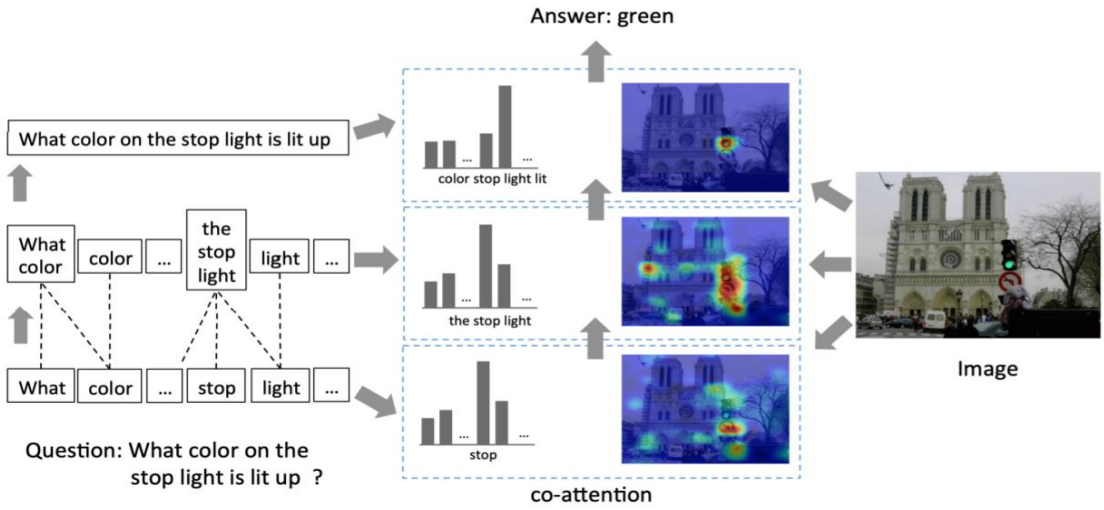

图 3. 问答系统应用注意力机制[34]

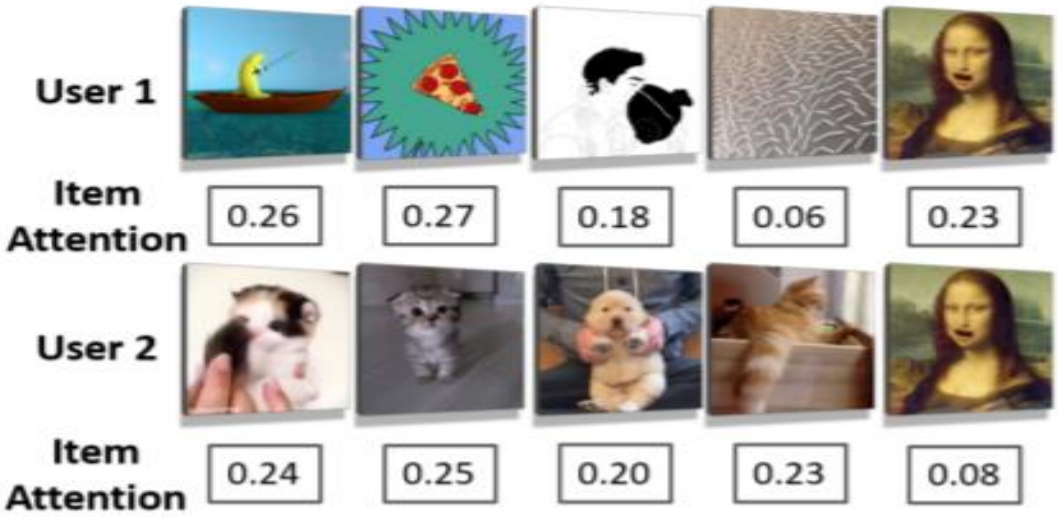

图 4. 推荐系统中应用注意力机制[35]

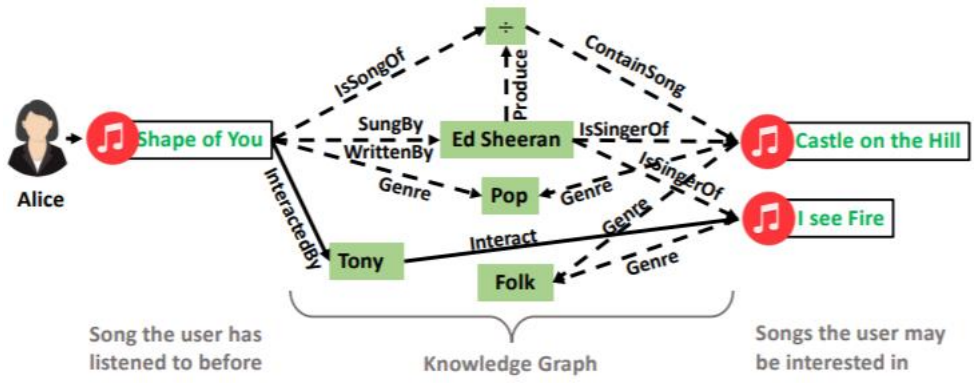

图 5 知识图谱在推荐系统中应用

2.2 近似模型

在学习模型完成学习后，利用可解释性方法或者构建与学习模型近似的解释模型，它的重点在于设计高保真的解释方法或者高精度的解释模型。细分为全局可解释性和局部可解释性。

2.2.1 局部可解释性

局部可解释性是以实例为目标，分析输入样本的每一维的特征变化[9]对输出结果的影响来代表整个模型的一个决策过程。Ribeiro M T[4]提出 LIME，它由两部分组成 LIME 和 SP-LIME,而 LIME 用一种保真的方法近似估计模型，SP-LIME 来挑选无冗余的实例（基本覆盖所有特征）进而解释模型的全局行为。LIME 除了能够对图像的分类结果进行解释外，还可以应用到自然语言处理的相关任务中，如主题分类、词性标注等。因为 LIME 本身的出发点就是模型无关的，具有广泛的适用性。但是在速度上却远远不如 Grad-CAM[30]那些方法来的快，Grad-CAM 通过分析梯度参数来进行可视化分析，与 CAM 相比，不用考虑模型结构的问题，因此避免了解释度与模型准确度之间的衡量。LIME 的速度依赖于模型本身的复杂度以及抽样数量，在选择完实例后需要对是每个实例进行抽样。

在 18 年新提出了 Anchors[31]的方法，指的是复杂模型在局部所呈现出来的很强的规则性的规律，Anchors 是对 LIME 的一种延续，对 LIME 缺点的补充。LIME 是在局部建立一个可理解的线性可分模型，并且可解释性的实例并不能完全代表整个模型的预测行为，对于高度非线性的模型，给出的解释并不一定能忠于模型本身。而 Anchors 是基于 if-else 建立一套更精细的规则系统。在和文本相关的任务上有不错的表现。LIME 不能很好的应用在 RNN 上，Grad-CAM 的方法适用于 CNN 为基础的模型上，因此 Guo W 等[2]提出了 LEMNA,这种方法适用于 RNN、MLP,支持局部非线性决策,并首次引入融合 lasso 来处理特征依赖性问题。

大多数局部解释方法是模型无关的，复杂的深度学习模型，模型的全局行为难以解释，对于特定输入做出一个准确的解释，让人能够理解这个黑匣子在特定输入上做出的预测，从而去理解整个模型。

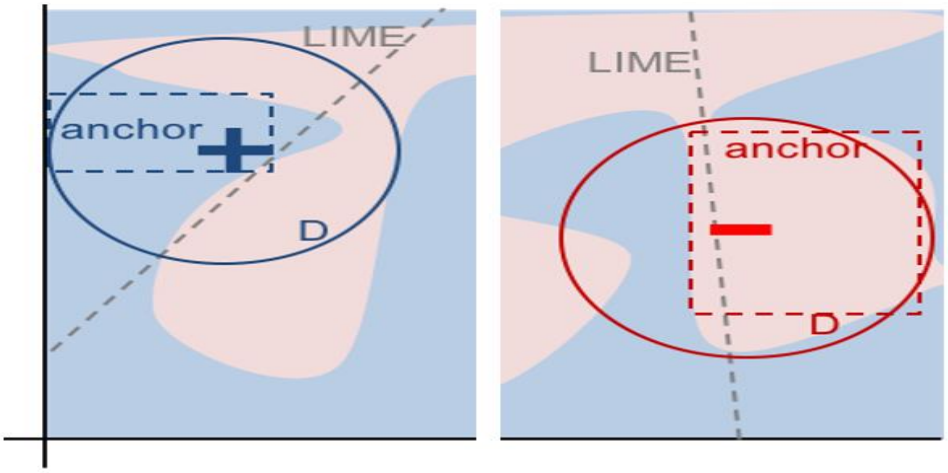

图 6 Anchor 与 LIME 在同一实例上的区别

图 7 Anchor 在图像中的应用，Anchor 在不同背景下的预测

2.2.2 全局可解释性

全局可解释性是指能够基于完整数据集上的依赖（响应）变量和独立（预测变量）特征之间的条件交互来解释和理解模型决策，理解模型的逻辑规则，以人可以理解的方式来表现一个复杂的模型。全局解释比局部解释更全面，因此在用于经过训练的模型或从模型中提取知识时更为有效。全局可解释性方法可以归纳为提取规则得方式和分析全局特征的方式。首先提取规则的方法中以树为基础，Deng H[28]提出的 inTree(可解释树)框架，可以应用于分类和回归问题，该框架从树集合中提取，度量，修剪和选择规则，并计算频繁的变量交互作用。也可以形成一个基于规则的学习器，称为简化树集成学习器（STEL），并将其用于将来的预测。Yang C 等[21]使用紧凑的二叉树（解释树）来明确表示黑盒机器学习

模型中隐含的最重要的决策规则，可以从单个数据样本的局部解释中全局性地解释黑盒机器学习模型，这是一个分割的方法，分割的置信区间难以确认。另一种基于规则的是用 if-else 来约束，PuriN[22]提出一种方法，该方法学习 if-then 规则以全局解释用于解决分类问题的黑匣子机器学习模型的行为。它的工作原理是首先提取在实例级别上很重要的条件，然后通过具有合适适应性函数的遗传算法发展规则。这些规则共同代表了模型进行决策的模式，对于理解其行为很有用。Ibrahim M 等[29]提出了一种称为 GAM 的全局属性生成方法，该方法解释了跨亚群的神经网络预测的格局，GAM 的全局解释描述了由神经网络学到的非线性表示。GAM 还提供可调整的子种群粒度以及跟踪对特定样本的全局解释的能力。Guo W, Huang S[3]等提出一个从目标模型提出概括性见解的方法，将多个弹性网引入贝叶斯非参数模型回归模型，利用该模型对目标模型进行近似，从而得出对模型概括性见解和针对单个决策的解释。相比前面的方法，Liu X, Wang X[7]提出的方法不会损害模型本身的精确度。尽管上述解释方法是根据原始模型直接开发的，但是缺乏对可解释性定量判断的标准。并且在提取一般规则时，我们会丢失一些细节。我们不知道这些细节对于高度可预测性有多重要。全局解释还有另一种重要的形式是自编码网络（AE），自编码网络能够学习到数据的重要特征，它的变体能够将特征解耦并表示，自编码网络由编码器和解码器组成，使得输入数据与输出数据尽可能地相同，提取基于整个数据的重要特征，进而达到解释的目的。

2.3 智慧医疗可解释性方法

上述提出的可解释性方法大多是针对图像和自然语言处理方面，一些通用的方法数据可视化、特征探索等仍然可以应用于医疗领域。例如通过图像识别的方法来识别 CT 图，细胞检验等。医疗信息拥有其独特的特征，在应用深度学习方法来解决医疗问题的同时，不可忽略这些重要信息，ECG 数据的时序性，每个人病患的独一性等[34,38-39]。

在医疗领域深度学习的可解释性大致分为两类，一类是以注意力机制为主，通过分析注意力机制的参数来解释特征，并通过实验证明添加注意力机制的模型的表现力和可解释性得到了提升。另一类是以无监督学习的自编码器（AE）以及其变种提取特征从而达到可解释性的目的。Van Steenkiste 等人[39]提出使用 VAE

产生可解释的 ECG 节拍空间的方法，能够从原始的 ECG 心拍中学习到代表心拍，并由此组成一个心拍空间，并表明一个心跳周期都可由这些基础心拍组成，从而达到对模型解释的目的。这两种方法都是基于模型本身的解释方法，解释更忠实于模型，并且没有损害模型的精确度，在精确度和可解释性之间得到了很好的平衡。

三．评价标准

可解释性的定义由行业内专家所给出，同时也提出了很多可解释性方法[15]，这些方法都有各自的优点和缺点，不同的研究领域使用不同的指标来衡量，但仍需一般性评价指标。如何用一个统一的标准去衡量？将从研究领域、目标用户和评价指标这三个方面来衡量可解释性方法。

3.1 研究领域

不同的研究领域对可解释性的要求是不同的，例如医疗，自动驾驶，网络安全[2]，金融领域，他们根据模型的预测做出的结果是会产生重大的影响，而像机器翻译，商品推荐等，容错性高，即使出现错误，也是可以包容的，那么提供可解释会给用户一个更好的体验，但相对来说可解释性要求没有前者高

3.2 目标用户

可解释性最终要以我们能够理解的方式呈现，呈现的方式需要根据不同的人做出区分。解释类型、长度、以及详细长度会受到解释目的和对象的影响。例如，机器学习专家希望通过解释方法，详细说明模型的细节或者是不足，并据此来优化和调整模型，以达到更好的效果。而医生在使用机器学习的方法辅助诊断时，只需要了解为什么会有这个诊断结果，预测结果是否能帮助医生找到遗漏的病症或是验证医生自己诊断的正确性。因此，用户的需求以及和专业水平对评估标准有较大影响。

3.2.1 AI 终端用户

AI 应用的终端使用者，他们在日常生活中使用 AI 产品（智能家居、社交网络、网购平台等）但不具备 AI 的专业知识，他们不需要了解模型内部详细的推理逻辑和决策过程，因此他们需要一个直观的，用户友好交互的一种解释方式，增强用户的体验感和信任度。在医疗领域，例如可穿戴设备和远程辅助诊断设备，当这些设备检测到用户的心电数据异常时，提醒用户及时就诊，但同时需要给出

用户解释，依据什么判断发现心电数据的异常，用户根据预测结果决定是否前往医院就诊，引起用户注意，从而减少因错过治疗时间导致的严重后果。

3.2.2 数据专家

数据专家中通常是使用机器学习来对数据进行分析、决策等各领域的专家，还包括一些研究人员。这类用户通常缺乏人工智能或机器学习算法的技术细节方面的专业知识，他们通常使用交互式数据分析工具、推荐系统或结合了交互式界面和算法的可视化分析系统，可解释性帮助他们判定人机任务表现。在医疗领域种，这一类用户通常是医护人员，他们通过机器学习模型给出的结果，为自己的诊断查漏补缺，使用心电数据诊断心脏问题的重要手段之一，医生通过心电图来做出判断，但是人眼的鉴别可能会遗漏一下不起眼的问题，或者是会发生错判，优秀的机器学习模型能够察觉这些细微的变法，为医生的诊断起到补充作用。因此，相对于 AI 的终端用户，他们是更需要机器学习模型的可解释性。了解模型的决策过程，发现数据中的偏差和错误，纠正自己的判断，对模型进行调整，提高就诊效率和速度，为病人提供可靠快速的就诊方案。

3.2.3 机器学习专家

机器学习专家是设计可解释性的机器学习算法以及其他机器学习算法的科学家和工程师。各种可视化和可视化分析工具帮助这些机器学习专家验证模型的准确性，提高可解释性和可靠性。他们需要准确的数值来衡量可解释性模型和可解释性方法，包括可解释模型的准确度，解释方法的保真度，生成解释速度，解释的一致性等，通过这些数值来发现模型的缺陷，例如模型敏感性，数据敏感性等，从而优化模型，提出忠实任务本身且全面可理解的机器学习模型，完成人工智能的落地。

3.3 评价指标

1)解释满意度

这个评价指标通常是对 AI 应用的终端用户，一般可以通过调查问卷或者是与用户的界面交互来完成，带有用户的主观判断，通过统计数据来评价解释的满意程度。[30]通过使用不同的解释方法，对模型可视化后，抽取 54 名工人进行主观判断解释的认可度，是否以人能够的理解的方式进行解释。

2)可解释模型的准确度和召回度

机器学习模型的复杂度一定程度上决定了模型的准确度，往往越复杂的模型能学到数据更多的特征，但同时模型的可解释性则会降低，而贝叶斯模型、决策树模型，可解释度高，但要牺牲模型的准确度。

3)可解释性方法的保真度

保真度是衡量可解释性方法是否能以最贴近的方式对模型做出解释，Guo 等人[2]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。Ribeiro M T[4]，随机假设 25%为不可信特征，并通过对逐步移除选择实例的不可信特征，观察预测结果的改变，如果发生改变则说明解释方法的保真度低，反之则高。[30]通过对输入数据的遮挡，解释方法给出了一个评估分数，通过分数的变化来判定该方法是否忠于模型。

4)可解释性方法的适用性

第二节总结了现存得一些可解释性方法，大多数得可解释性方法应用于计算机视觉和自然语言处理领域。而自动驾驶、智慧医疗、网络安全等更需要可解释性，现存的方法并不是十分适合。根据行业特征要调整可解释性方法或者是通用的解释方法，这是一个衡量的重要因素。

四．总结

随着人工智能的发展，机器学习的可解释性是一个必须的过程，理解机器学习内部的逻辑结构和运作机制，有一个全面而准确的认识，才能在人与人工智能构建友好的沟通机制，才能依据机器的预测结果帮助我们做出正确的决策。可解释性方法是多种多样的，需要根据数据本身特征以及任务要求选择合适的可解释性方法，达到理解模型和提高模型性能的目的。一般解释方法在医疗领域有一定的局限性，医疗领域的高安全性要求必须对深度学习模型给出准确而全面的解释，才能是深度学习更好的应用在临床中。未来我们需要更加精确的评价指标来衡量可解释性方法，局部的可解释方法更加贴近模型本身，而全局的解释方法有可能会忽略模型中一些信息，可以将两者的优点相结合，在不影响模型精度的情况下，给出更合理且全面的解释方法。

五．参考文献

[1] 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[J]. 计算机研究与发展, 2019, 56(10): 2071-2096. Ji Shouling, Li Jinfeng, Du Tianyu, Li Bo. Survey on Techniques, Applications and Security of Machine Learning Interpretability. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.

[2] Guo W, Mu D, Xu J, et al. Lemna: Explaining deep learning based security applications[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2018: 364-379.

[3] Guo W, Huang S, Tao Y, et al. Explaining Deep Learning Models--A Bayesian Non-parametric Approach[C]//Advances in Neural Information Processing Systems. 2018: 4514-4524.

[4] Ribeiro M T, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016: 1135-1144.

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

[6] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

[7] Liu X, Wang X, Matwin S. Interpretable deep convolutional neural networks via meta-learning[C]//2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-9.

[8] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1885-1894.

[9] Tamagnini P, Krause J, Dasgupta A, et al. Interpreting black-box classifiers using instance-level visual explanations[C]//Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics. ACM, 2017: 6.

[10] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Advances in Neural Information Processing Systems. 2017: 4765-4774.

[11] Wu M, Hughes M C, Parbhoo S, et al. Beyond sparsity: Tree regularization of deep models for interpretability[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[12] Wu C, Gales M J F, Ragni A, et al. Improving interpretability and regularization in deep learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 26(2): 256-265.

[13] Fong R C, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3429-3437.

[14] 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.

[15] Mohseni S, Zarei N, Ragan E D. A survey of evaluation methods and measures for interpretable machine learning[J]. arXiv preprint arXiv:1811.11839, 2018.

[16] Hicks S, Riegler M, Pogorelov K, et al. Dissecting Deep Neural Networks for Better Medical Image Classification and Classification Understanding[C]//2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2018: 363-368.

[17] Chua Z L, Shen S, Saxena P, et al. Neural nets can learn function type signatures from binaries[C]//26th {USENIX} Security Symposium ({USENIX} Security 17). 2017: 99-116.

[18] Krusinga R, Shah S, Zwicker M, et al. Understanding the (un) interpretability of natural image distributions using generative models[J]. arXiv preprint arXiv:1901.01499, 2019.

[19] Alvarez-Melis D, Jaakkola T S. Towards robust interpretability with self-explaining neural networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., 2018: 7786-

7795.

[20] Seo S, Huang J, Yang H, et al. Interpretable convolutional neural networks with dual local and global attention for review rating prediction[C]//Proceedings of the Eleventh ACM Conference on Recommender Systems. ACM, 2017: 297-305.

[21] Yang C, Rangarajan A, Ranka S. Global model interpretation via recursive partitioning[C]//2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018: 1563-1570.

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

[23] Wang J, Gou L, Zhang W, et al. DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation[J]. IEEE transactions on visualization and computer graphics, 2019, 25(6): 2168-2180.

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

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

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

[27] Liu X, Wang X, Matwin S. Improving the Interpretability of Deep Neural Networks with Knowledge Distillation[C]//2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018: 905-912.

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

[29] 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.

[30] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.

[31] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[32] Chaudhari S, Polatkan G, Ramanath R, et al. An attentive survey of attention models[J]. arXiv preprint arXiv:1904.02874, 2019.

[33] Choi E, Bahadori M T, Sun J, et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism[C]//Advances in Neural Information Processing Systems. 2016: 3504-3512.

[34] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]//Advances In Neural Information Processing Systems. 2016: 289-297.

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

[36] Wang X, Wang D, Xu C, et al. Explainable reasoning over knowledge graphs for recommendation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 5329-5336.

[37] Bai T, Zhang S, Egleston B L, et al. Interpretable representation learning for healthcare via capturing disease progression through time[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 43-51.

[38] Ming Y, Xu P, Qu H, et al. Interpretable and Steerable Sequence Learning via Prototypes[J]. 2019.

[39] Van Steenkiste T, Deschrijver D, Dhaene T. Generating an Explainable ECG Beat Space With Variational Auto-Encoders[J]. arXiv preprint arXiv:1911.04898, 2019.

[40] Li R, Zhang X, Dai H, et al. Interpretability Analysis of Heartbeat Classification Based on Heartbeat Activity's Global Sequence Features and BiLSTM-Attention Neural Network[J]. IEEE Access, 2019, 7: 109870-109883.

[41]胡盛寿，高润霖，刘力生，等. 《 中国心血管病报告 2018》 概要[J]. 中国循环杂志，2019 (3)：2.