# 深度学习真的是黑盒吗？可视化和可解释性

随着深度学习的发展，深度学习在各个应用领域上取得了其他传统算法所无法比拟的极高精准度，越来越多的研究者在自己的研究领域中应用深度学习算法；许多商业公司使用深度学习算法提高自身竞争力；就连许多政府决策都尝试利用深度学习算法提升效率。但是，相比于深度学习技术在各个领域应用的高速扩展，我们对于深度学习技术的理解其实不甚了解。深度学习模型内部复杂的网络连接，使得我们很难了解模型内部究竟发生了什么。因此，深度学习模型被认为是一种“黑盒”。但是深度学习的黑盒性并不是来源于封装，而是来源于复杂性：简单的ANN也容易被人理解，复杂的决策树也一样令人费解。深度学习系统并不是对外完全不可见的，而是缺少一种简单的、易于理解的、可信的理解方式，让人们可以理解其中的原理，进而取得用户的信任。研究者们尝试使用各种方法说明深度学习系统究竟在学习过程中发现了什么特定模式，基于这些研究成果，研究们提出了一种新的人工智能范式——可解释人工智能（Explainable Artificial Intelligence，XAI）。

## 为什么需要解释深度学习

为什么我们需要解释深度学习？针对这个问题，我们将从一个例子出发来进行分析。

在一次进行图像分类的任务中，使用了若干张哈士奇和狼的样本图片，如图14-1（a）和14-1（b）所示。经过一个简单的深度学习模型的训练过程，在大多数的数据中都能够正确的分类，但是发现一些类似图14-1（c）的图片总是被错误分类。经过可解释性的研究，人们发现模型对图14-1（c）分类的依据如图14-1（d）所示。也就是说，之所以图片被误判，是因为输入数据中，所有哈士奇都在草地上，所有狼都在雪地上，深度学习模型误以为可以通过“草地/雪地”来区分“哈士奇/狼”。当人类使用者了解到这样的判决情况后，尽管这个模型的判别准确率高达90%以上，但是使用者对模型的信用度下降到11%。[1]



图 14‑1 深度学习模型的可解释性

从这个例子，我们可以看出，目前为止，基于数据驱动的人工智能系统并不能完全获得人类的信任，深度学习模型很容易被“愚弄”而导致模型的可靠性和泛化性变差，这也限制了人工智能在各种领域的应用。

### 深度学习决策机制的理论局限

图14-1所展示的例子以一种具象的形式向我们说明了大多数深度学习模型的局限性。如果模型仅仅只能在输入数据和预期结果之间建立关联（Association），由于数据样本自身的局限和偏差，这种关联学习将会学习到虚假关联（Spurious Relationship），例如狼与雪地之间的相关性。这样的模型在训练数据分布和测试数据分布相近的情况下表现良好，但其实并没有学习到基于正确因果关系的推理决策能力，在训练和测试数据分布不一致的情况下，泛化性能就会变差。

Judea Pearl提出，基于关联分析的学习方式是一种低层次的认知，而为了能够从这些虚假关联中甄别出因果关系，需要通过主动干预（Intervention）实验来拓展观测现象，利用反事实推理（Counterfactual Reasoning）发现其中的内在因果联系。[2]这也揭示了人工智能和深度学习的未来发展方向之一。

### 深度学习模型的应用缺陷

在深度学习模型的实际应用过程中，数据驱动的深度学习同样具有很多隐患。

首先，数据样本的局限和偏差会导致训练得到的模型同样具有偏差。例如我们在图14-1说明的例子。又比如，在美国曾经使用过的犯罪风险评估算法COMPAS因为训练数据的偏差，被发现黑人相比于白人更容易被判定为高犯罪风险。这种具有偏差的人工智能系统既不能获得人们的信任，还不利于社会的公平正义。

其次，当前深度学习模型容易受到攻击。如图14-2所示，在一张被模型分类为熊猫的图片上，添加人眼不可见的特定噪声，得到的新图片会被同样的模型分类为长臂猿，这样的攻击被称为对抗学习。[3]利用这种特性，人们可以通过佩戴经过特殊设计的眼镜[4]，就可以欺骗使用深度学习进行人脸识别的人脸识别系统，这会引发用户对系统的不信任。



图 14‑2 深度学习模型受到特定噪声攻击

从这些风险可以看出，现在的深度学习系统依旧缺乏对其工作过程的具体认知，难以解释决策流程，总体上难以获得人们的足够信任，尤其是在金融、法律、医疗等错误代价较高的领域，人们不能信任这样的深度学习系统。

### 深度学习模型面临的监管要求

事实上，目前在金融、医疗和法律等重要领域，政府对深度学习系统已经出台了许多防范措施和监管立法。比如，欧盟高级人工智能专家组起草的《可信人工智能道德原则指导》[引用来源：Ethics guidelines for trustworthy AI...]指出，可信人工智能系统必须满足七个方面的要求:人类监管纠错、技术安全及鲁棒、隐私保护和数据治理、透明及可解释、算法公平及无歧视、环保及社会影响、问责制度。中国新一代人工智能治理专业委员会在2019年6月发布的《新一代人工智能治理原则——发展负责任的人工智能》[引用来源：发展负责任的人工智能...]政策文件中指出，要突出发展负责任的人工智能，强调公平公正、尊重隐私、安全可控等八条原则，因此，提高可解释性成为推广智能模型在各行各业应用的必由之路。

虽然对人工智能可解释性的监管要求已经在法律和规章制度层面逐步完善，但如何将这些制度层面的规则具体细化落实为可实现的技术方案，仍是可解释人工智能巫待研究和解决的挑战。

### 解释与信任

通过前面对深度学习局限性的解释，我们有这样的认识：解释深度学习的本质要求就在于通过解释的方式取得人类对深度学习系统的信任。信任（Trust）是人类在社会协作中的一种心理状态，是我们愿意向对方暴露自己的弱势。我们信任深度学习系统，建立在两个层次的主观判断，其一是对深度学习系统能力和性能的信任，信任系统有处理问题的能力；其二是对深度学习系统态度和决策机制的信任，信任系统能够做出合乎人类道德规范的判断决策。[5]

解释（Explanation）是人类相互沟通并获取信任的基本方法。这一点同样适用于各种智能体之间，深度学习系统要获取人类的信任，同样需要通过解释和沟通进行“磨合”。而类似于人与人之间的解释过程，深度学习系统的解释过程同样要考虑到受众——主要是使用者的需求，使用人类可以理解的形式进行解释，从而获取人类的信任。[5]

## 可解释深度学习

### 定义

可解释深度学习需要在一种人机沟通的交互场景中，逐步学习和调整其解释，并最终完成任务。在这个深度学习系统中，除了作为决策主体的深度学习模型和使用主体的使用受众，还需要一个提供解释的解释者，这个解释者可以由模型本身充任，也可以是第三方的解释系统或者人类专家。解释者要根据当前的任务场景、任务和受众的知识水平等因素选择合适的解释方法，并在调整优化的过程中优化解释，来取得人类受众的信任。

### 解释结果的评价标准

解释的结果要考虑到受众的知识水平等因素，对于相同任务相同模型的不同解释结果和解释形式，人类受众必然具有一定的偏好，例如人们总是倾向于简单的、可视的和可靠的解释结果。因此对于解释结果人们形成了一些评价标准。[5]

(1)可解释性方法的敏感度。敏感度是指可解释性方法在受到噪声干扰时，提供的解释结果是否会对噪声敏感。此类可解释性方法可进一步细分为对模型参数加噪声与对输入样本加噪声。一般来说，可解释性方法对于模型参数上的噪声敏感度需要相对较高，这表明可解释性方法与模型参数高度相关，而对输入样本上的噪声敏感度需要相对较低，这使得可解释性方法在有输人干扰的情况下也能给出可靠的解释。

(2)可解释性方法的对抗攻击鲁棒性。近年来，深度学习面临的一大挑战是对抗样本的存在，而类似的挑战也同样存在于可解释性领域。针对可解释性的对抗攻击在输人样本上添加特定的扰动，能够在不改变人工智能系统预测结果的情况下改变解释结果。因此，目前的一些评测算法希望量化可解释性方法的对抗攻击鲁棒性，这有助于人们找出能够抵御对抗攻击的可解释性方法，从而得到更为可靠的解释结果。

(3)可解释性方法的全面性。人们关注的可解释性方法的另一个性质是全面性，它反映了一个解释结果是否能完全反映人工智能系统全部的信息处理逻辑。某些可解释性方法只能够对系统的一部分进行解释，而另外一些方法则能够对系统整体做出解释。在这种情况下，后者的全面性要好于前者。

(4)可解释性方法的客观性。可解释性方法的客观性（在某些工作中被称为正确性或忠实度）是衡量一个可解释性方法的重要指标。所谓客观性，指的是可解释性方法是否客观地反映了人工智能系统的处理逻辑，比如解释结果告诉人们系统中的某个变量很重要，而这个变量对于人工智能系统确实很重要，那么这个解释就是客观的。对可解释性方法客观性进行评测

能够帮助人们选择真正可靠的可解释性方法。

(5)解释结果的简单易懂性。评测可解释性方法的另一个重要的角度是解释结果的简单易懂性，即解释结果本身要简单易懂，让人容易理解解释结果的简单易懂性主要体现在三个方面:首先，我们需要用少数概念解释模型内部的复杂逻辑，而不是用成千上万个大量概念进行解释；其次，我们需要保证这些概念本身的语义是人们能够理解的，而不是人们无法认知的特征；最后，简单易懂性与客观性有时是对立统一的，当一个模型本身的概念很复杂时，追求简单易懂性就变为对真实逻辑的近似，因此，人们需要做好二者的平衡。同时，让黑盒模型尽可能使用精炼可靠的特征表达进行预测才更容易使解释结果简单易懂。

(6)可解释性方法的互洽性。由于目前有越来越多的可解释性方法涌现这些方法之间是否互洽也是评测可解释性方法的一个指标，即不同解释性算法的解释结果需要相互印证，不能只是自圆其说。可解释性方法之间的互洽性可以体现在对某个变量重要性的不同解释结果之间相互印证上，也可以是对人工智能系统表达能力的解释，比如对神经网络迁移性或泛化性的不同解释，二者应该是殊途同归的，而不能相互矛盾。

(7)可解释性方法的计算效率。在实际应用中，人们往往也会关注可解释性方法的计算效率。一般来说，人们更希望能够在较短的时间内得到解释结果，然而一些可解释性方法的计算复杂度过高，比如需要NP-hard时间复杂度的Shapley value，以及一些通过训练进行解释的方法等，使得它们难以被投人对实时性要求较高的一些应用，如金融业的投资决策。

### 解释的分类

对于深度学习的解释形式多种多样，根据不同的引用场景和任务需要采用合适的解释形式。但是根据解释说明的目标，将解释结果大致分为两大类：局部解释（Local Explanation）和全局解释（Global Explanation）。

局部解释是解释者针对一个特定的输入实例，给出深度学习模型进行决策的依据。例如一个进行图像分类的深度学习模型，针对图14-3（a）所示的图片分类结果为“猫”，针对这个决策结果，解释者通过给出比较重要的图片区域的方法给出了局部解释，如图14-3（b），说明了模型“为什么觉得这是猫”。

全局解释是解释者针对一类特定的输出结果或者一系列相近的输出结果，给出模型认为的共同特征或者典型特征。例如图14-3（c）的图片展示了，对于类别“猫”，解释器给出了模型认为的类别的典型特征，作为“全局解释”，说明了模型“觉得猫应该长成什么样”。



图 14‑3 深度学习模型的局部解释和全局解释

两种解释都在一定程度上刻画了模型的解释过程，有助于受众理解模型的决策过程，而具体要采用哪种解释结果，则应该针对具体受众和任务特点确定。

### 解释方法的分类

解释方法是指解释者在解释模型的决策过程时所采用的方法。按照解释者和深度学习模型的关系可以分为事前可解释建模方法（Ante-hoc Explanation）和事后可解释分析方法（Post-hoc Explanations）。[6]

事前可解释建模方法，是指在模型训练之前，将可以提高深度模型的可解释性的方法结合到模型结构中来实现可解释性，例如使用特殊设计的正则化方法来约束模型的训练以提高可解释性[7]；添加注意力机制来提高可解释性[8]–[10]；也有研究者将贝叶斯网络或因果学习等方法结合进深度学习模型中来提高可解释性[11], [12]。事前可解释建模方法依赖于模型自身具备的可解释性，而模型的可解释性和模型的预测性能往往不可兼得，需要研究者进行实验和调参来权衡可解释性和模型的预测性能。[13]

事后可解释分析方法是在充分学习模型之后进行的，旨在利用事后解释方法或构建解释模型去解释模型的工作机制或决策依据。事后解释的方法具有很多优势，由于预测和解释是两个独立的过程而不会相互干扰，因此不需要为了追求可解释性来牺牲预测性能。[6]此外，由于事后可解释分析方法较少的参与建模过程，因此事后可解释分析方法大多模型无关（Model Agnostic），使得这些方法的引用更加广泛。但是，事后解释通常并不完全忠实于原始模型，我们使用可解释性方法是为了以更加简单清晰的方法展示原始模型的决策过程，这个过程中必然伴随对原始模型的一些细节的忽略。因此，事后解释方法或多或少都是不准确的，也导致我们无法完全信任这些方法。

本章接下来的部分将主要集中介绍事后可解释分析方法。其他事前可解释建模的方法请读者参阅其他章节和相关文献。

## 网络特征可视化

### 最大激活响应可视化

### 基于反向传播的输入重建可视化

## 类别相关映射可视化

## 基于扰动的解释方法

## 利用梯度的解释方法

### 显著性图

### 积分梯度

## 逐层相关性传播

## 基于代理模型的方法

## 其他解释性方法

## 小结

## 参考文献

[1] M. T. Ribeiro, S. Singh和C. Guestrin, 《〈Why Should I Trust You?〉: Explaining the Predictions of Any Classifier》. arXiv, 2016年2月16日. 见于: 2022年10月9日. [在线]. 载于: http://arxiv.org/abs/1602.04938

[2] J. Pearl和D. Mackenzie, *The book of why: The new science of cause and effect*. Penguin (May 2, 2019).

[3] I. J. Goodfellow, J. Shlens和C. Szegedy, 《Explaining and Harnessing Adversarial Examples》. arXiv, 2015年3月20日. 见于: 2023年5月5日. [在线]. 载于: http://arxiv.org/abs/1412.6572

[4] X. Chen, C. Liu, B. Li, K. Lu和D. Song, 《Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning》. arXiv, 2017年12月14日. 见于: 2023年5月5日. [在线]. 载于: http://arxiv.org/abs/1712.05526

[5] 杨强, 范力欣, 朱军, 陈一昕, 张拳石和朱松纯, 可解释人工智能导论, 1 本. 北京: 电子工业出版社.

[6] Z. C. Lipton, 《The Mythos of Model Interpretability》. arXiv, 2017年3月6日. 见于: 2023年5月6日. [在线]. 载于: http://arxiv.org/abs/1606.03490

[7] M. Wu等, 《Regional Tree Regularization for Interpretability in Black Box Models》. arXiv, 2020年3月16日. 见于: 2023年5月6日. [在线]. 载于: http://arxiv.org/abs/1908.04494

[8] A. Vaswani等, 《Attention Is All You Need》, 收入 *arXiv:1706.03762 [cs]*, 12月 2017. 见于: 2022年2月28日. [在线]. 载于: http://arxiv.org/abs/1706.03762

[9] A. Dosovitskiy等, 《An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale》. arXiv, 2021年6月3日. doi: 10.48550/arXiv.2010.11929.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò和Y. Bengio, 《Graph Attention Networks》, 收入 *arXiv:1710.10903 [cs, stat]*, 2月 2018. 见于: 2022年2月25日. [在线]. 载于: http://arxiv.org/abs/1710.10903

[11] C. Li, M. Welling, J. Zhu和B. Zhang, 《Graphical Generative Adversarial Networks》. arXiv, 2018年12月12日. doi: 10.48550/arXiv.1804.03429.

[12] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He和Z. Shen, 《Deep Stable Learning for Out-Of-Distribution Generalization》. arXiv, 2021年4月15日. doi: 10.48550/arXiv.2104.07876.

[13] 邵平, 杨健颖和苏思达, 可解释机器学习：模型、方法与实践. 北京: 机械工业出版社.