**Neural network interpretability**

**一、研究动机**

为什么我们需要研究神经网络的可解释性？简言之，当AI被应用于某一领域，尤其是需要确保输出结果可靠性的领域，我们必须获得可解释的输出结果，以规避不确定性可能带来的麻烦和问题。

**（一）确保AI司法行政的公平性**

当AI被应用于司法行政辅助，尤其是刑事审判辅助时，我们需要确保AI系统的输出结果是公平的，为了我们需要对其输出结果进行明确的解释，例如假释结果判定、犯罪结果预测、警务派遣规划等。

**（二）确保金融决策的正确性**

当AI被应用于金融辅助，例如信用评分、贷款审批、保险报价时，我们不仅需要其预测结果，还需要AI对结果的产生进行理论的解释。

**（三）确保医疗卫生的安全性**

当AI被应用于医疗服务，我们必须确保其输出结果有理有据，并对患者的安全负责。

**（四）确保关键系统的可靠性**

当AI被应用于重大系统项目，如航天、军事等，必须依赖其可解释的确定结果，确保任务的成功率，为人类的前途命运负责。

总结而言，研究神经网络的可解释性可以解决对AI的安全担忧和伦理法律纠纷的问题，提高社会对AI模型的认可度，并且这也将有助于模型性能的调试，加深我们对于神经网络原理的理解。

**二、可解释性的分类**

**（一）本质的、被动的（事后的）、主动的**

本质可解释性意味着模型本身就可以直接解释，通常用于简单的模型，如单层感知器。

被动（事后）可解释性是指在模型训练后，独立于模型而使用的各种解释方法。

主动可解释性是指主动改变网络架构或训练过程，以获得更好的可解释性。

**（二）根据解释方法的输出进行分类**

示例：提供相似的或原型的示例。

归因：为输入特征（例如特征重要性、显著性掩码）分配信用（或指责）。

隐藏语义：理解某些隐藏的神经元/层。

规则：用可解释的模型（如决策树、规则集和其他规则格式）近似黑盒模型。

**（三）依赖于模型、独立于模型**

依赖于模型：例如，特征图可视化只适用于神经网络模型。

独立于模型：输入和输出的关系分析可以应用于几乎所有的预测模型，如特征降维、主成分分析等。

**（四）局部的、全局的**

局部的：解释网络对单个样本的预测（例如输入图像的显著性掩码）。

半局部的：解释一组相似的输入。

全局的：解释整个网络（例如一组规则/决策树）。

**三、如何评估可解释性**

可解释性是一个潜在的属性，我们无法直接测量，因此我们依赖于可测量的结果对可解释性进行评估，而没有一定的基准。

一个好的解释是怎样的？它应该注重三个方面：（1）对比，一个好的模型解释要强调其与其他模型的不同；（2）选择，对模型的解释应注重其最重要的部分；（3）异常，对模型的解释应额外关注其产生异常值的情况。

我们期望解释包含以下特性：信息性、低认知负担、可用性、忠诚性、稳健性、无误导性、交互、算法复杂度。

在实践中，可解释性的评估方法有：

（1）应用：可解释性在多大程度上能有助于最终任务（例如，更好地识别错误或减少歧视）。

（2）人工：直接让人们主观评价解释的质量（例如，让用户在几种解释中选择哪种解释质量最高）。

（3）功能：使用可解释性的一些形式定义作为解释质量（例如稀疏性）的代理。

**四、DNN的可解释性方法**

**（一）利用规则解释**

逻辑规则、对比解释法（CEM）、反事实的视觉解释、关键数据路由路径（CDPR）、混合规则集（HyRS）、深度模型的树正则化。

**（二）利用隐藏语义解释**

激活最大化、网络剖析、相关性分析、可解释的CNN。

**（三）利用归因解释**

基于梯度的方法、引导反向传播、引导梯度CAM（类激活映射）、LIME、ExpO、DAPr、Dual-net。

**（四）利用示例解释**

通过影响函数、基于原型的案例推理、ProtoPNet