# 归因可解释评估平台

## 系统背景

作为近些年人工智能领域最流行的研究趋势之一，深度学习已经成为当前人工智能领域最为成功的学习范式。其中，以深度神经网络为代表的深度模型因其强大的特征学习能力和并行性高的特点，已经被广泛应用于各种领域。但随着深度学习技术的快速发展，深度神经网络模型的“黑盒”特性带来的不足越发明显。高度耦合的模型导致模型内部的决策过程难以解释，限制了其在医疗诊断、欺诈交易检测等对可解释性要求较高的场景中的应用。

因此，深度模型的可解释性学习已成为人工智能领域的一个重要研究方向，并催生了众多相关方法。其中，归因分析在深度学习中引起了广泛关注。近年来，研究人员从不同角度提出了大量归因方法，主要包括基于梯度、基于扰动和基于CAM的方法，如何有效且全面地评估这些可解释方法尤为重要。然而，目前的这些工作使用不同的数据集、模型和评估协议，没有一个有效且全面的评估指标。与此同时，目前的工作大多停留在科学研究的阶段，算法的使用需要较强的专业知识，缺乏针对面向业务执行人员的可解释评估平台，给非专业人员的使用带来了不便。

本项目针对基于归因分析的可解释学习方法，构建了用户友好的模型可解释评估平台Meta-Rank，通过聚合了来自不同评估配置的排名结果，减轻了与不同配置之间的偏差，实现了对可解释方法的全面评估，评估结果可视化以及可解释方法的综合分析。此外提出了一个评估指标，以促进更全面的基准比较。本项目通过友好的可视化界面降低平台使用难度，使其不仅适用于专业人员，也方便非专业人员的评估与使用。

## 主要功能

设计与实现基于归因分析的可解释评估平台，具体目标分为以下两个方面：

（1） 通过建立一个全面而统一的可解释评估指标，确保对可解释方法评估的公平性，实现可信的评估方案。

（2） 实现对用户友好的可视化界面以及可解释评估的算法接口，降低平台的使用难度，使其可解释评估的使用更加方便、快捷。

## 开发环境

开发环境：Linux

编程语言：python、matlab等

编程框架：tensorflow、pytorch、web开发框架

## 技术路线

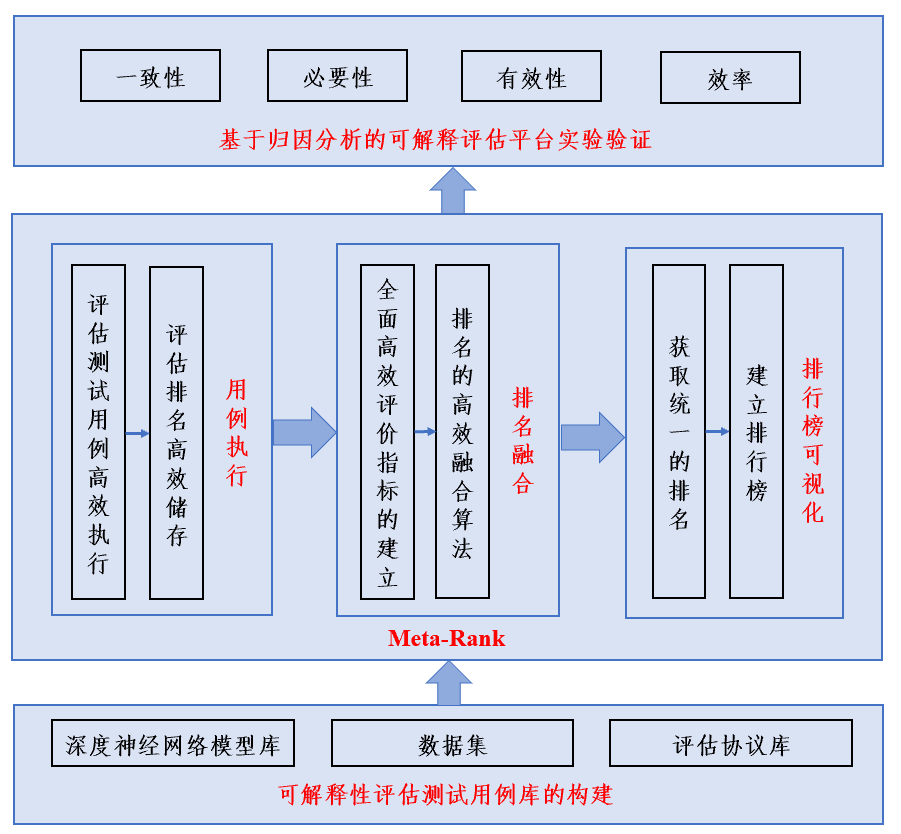


图 1 可解释评估平台meta-rank

关键技术：

归因分析可解释学习、排名融合算法、平台开发

技术难点：

实现对海量深度学习模型、数据集、评估协议的高效储存与运用；

对多个现有可解释评估排名的高效融合算法，实现统一而全面地排行；

设计友好的可解释评估接口以及排行榜可视化方法，提升平台的使用效率，降低平台使用的难度。