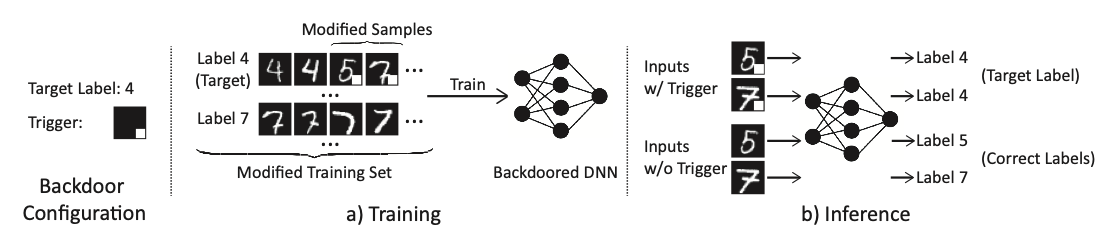
1. **研究人工智能系统的数据可靠性分析。**

研究对于训练数据选取的方法，根据模型对于数据的取值范围给出明确的定义。基于基于概率统计分析研究可靠性分析的方法，对选取的数据进行调整，从而保证模型准确率。研究训练数据的权重分布，对不均衡数据设计权重调整方案。针对被污染数据导致的后门问题，研究被污染数据的检测方案。通过建立深度学习后门检测系统，对现有的模型进行检测。结合以探索的后门触发方式，对输入数据进行深度检测。

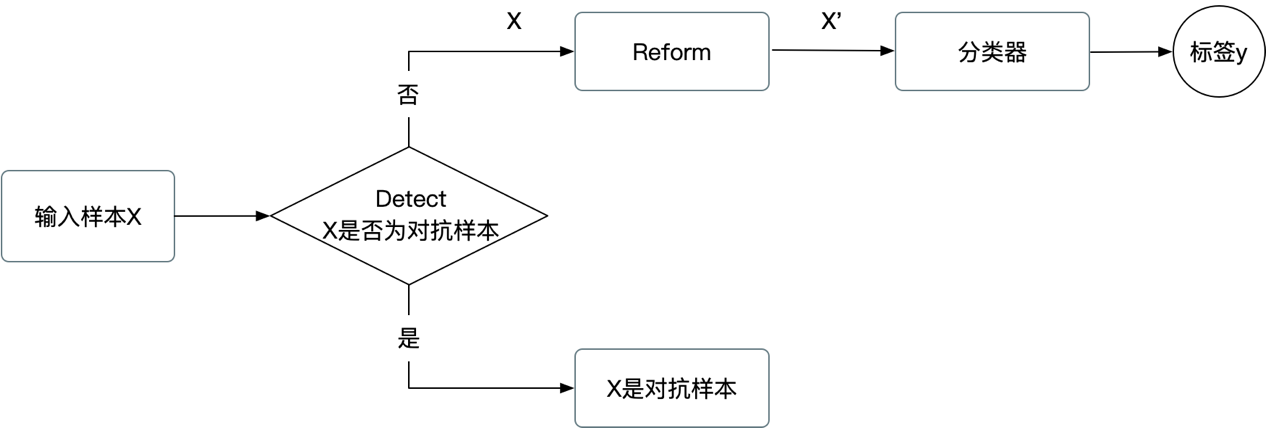


针对现有模型中输入数据的细微改变导致的错误分类问题，构建对抗样本检测系统。结合现有对抗样本生成方式，获得对抗样本数据特征，训练对抗样本识别工具，对输入模型的数据进行检测和变换，保证实际应用中数据可靠性。针对神经网络的黑盒特性，难以构建基于神经网络的无错误系统。利用探针技术，构建模型状态识别系统，获得深度学习模型运行状态。结合分层模型构建的状态转换图，设计状态匹配算法，对模型运行时状态转换和合理性进行判断。

1. **数据可靠性分析技术**

针对训练数据属性合理取值，数据分布不均匀的问题，利用让样本在训练过程中拥有相同的话语权的思想，可以利用模型融合的方法：多次下采样或者可以利用增量训练的思想。最终组合所有分类器的结果得到最终结果。针对系统结构中数据安全性检测问题，通过测量将所有输入从每个区域更改为目标区域所需的最小扰动量检测这些后门所处的位置。确定后门对应的触发器，然后在数据输入模型之前进行判断，将带有触发器的数据过滤。对于已经发现含有后门的模型对模型本身进行修补。

针对应用场景中数据可靠性分析，将对抗样本的识别问题分解为两个部分。通过分析对抗样本的生成算法，设计针对该算法的对抗样本检测系统。对于明显具有对抗样本特征的数据直接过滤；对于不能确定的数据找到相似的数据再利用分类器进行判断。



针对模型状态空间数据可靠性研究，在一个经过训练的模型的内部的每个层中添加探针，通过差异估计来验证模型的内部状态，根据模型层次分析构建模型的DFA图，根据运行时状态与DFA的匹配模式确定结果的可信度。

评测指标

对于数据的可靠性，要形成数据训练时检测，输入时检测，运行时检测的模型，并且对于模型中隐藏的后门提出有效的检测手段。