

智能软件可靠性综合评测指标体系模型的构建

高德鹏^{1*}

1. 南通理工学院, 用友数智学院, 江苏 南通, 226000

摘要: 人工智能驱动下的智能软件应用规模持续扩大, 传统软件的可靠性评测方法因难以适配数据驱动的深度神经网络模型, 无法直接用于智能软件的可靠性评估。针对目前智能软件缺少系统化、完备化可靠性评测指标体系的问题, 本文从模型、数据、运行平台三个维度解析智能软件可靠性的内涵, 面向目标识别、目标检测、智能决策典型业务场景, 构建包含通用性能评测指标、智能计算能力评测指标、智能任务评测指标、可靠性评估指标的综合评测指标体系模型。该体系从平台资源消耗、智能计算加速效能、智能任务输出效果、软件运行失效等多个层面实现对智能软件可靠性的多维度度量, 可为智能软件的可靠性测试、性能优化与质量评估提供理论参考与评测依据。

关键词: 智能软件; 深度学习; 可靠性; 指标体系; 综合评测

Construction of Comprehensive Evaluation Index System Model for Intelligent Software Reliability

Depeng Gao^{1*}

1. School of Yonyou Digital Intelligence, Nantong Institute of Technology, Jiangsu Nantong, 226000, China

Abstract: Driven by artificial intelligence, the application scale of intelligent software keeps expanding. Traditional software reliability evaluation methods cannot be directly applied to assess the reliability of intelligent software, as they are incompatible with data-driven deep neural network models. Aiming at the lack of systematic and complete reliability evaluation indicator system for intelligent software at present, this paper analyzes the connotation of intelligent software reliability from three dimensions: model, data and operation platform. Oriented to typical business scenarios including target recognition, target detection and intelligent decision-making, a comprehensive evaluation indicator system model is constructed, which covers general performance evaluation indicators, intelligent computing capability evaluation indicators, intelligent task evaluation indicators and reliability evaluation indicators. The system realizes multi-dimensional measurement for the reliability of intelligent software from the perspectives of platform resource consumption, acceleration efficiency of intelligent computing, output effect of specific intelligent tasks, and software operation failure characteristics. It can provide theoretical references and evaluation basis for reliability testing, performance optimization and quality assessment of intelligent software.

Keywords: Intelligent software; Deep neural network; Reliability; Indicator system; Comprehensive evaluation;

智能软件是人工智能的重要载体, 承载着人工智能算法和实现, 主要包括: 为人工智能提供支持的软件基础设施和产生智能行为的应用软件。随着人工智能的发展, 智能软件也表现出蓬勃发展的趋势, 在社会各领域展现出广泛应用价值^[1], 与此同时, 智能软件面临着必不可少的可靠性测试问题。

传统软件的可靠性测试过程一般是基于代码逻辑制定相应的测试用例, 然后运行测试用例并对

软件的功能性和非功能性进行测试,技术相对较为成熟。然而,智能软件的核心是基于深度神经网络的智能算法,它是数据驱动型软件,其内部逻辑主要由神经元之间的连接权值所决定,所以传统软件的测试技术对于智能软件并不适用^[2-5]。

智能软件的性能与支撑深度神经网络运算的软硬件平台、训练数据、测试数据与训练数据之间的分布差异等内容相关,但是,智能软件的可靠性具体地反映在哪些指标上目前尚无定论。因此,为了能够全面地反映智能软件的可靠性,本文拟面向目标识别、目标检测、智能决策等典型应用场景构建智能软件可靠性综合评测的指标体系模型,主要包括以下几个方面:通用性能评测指标、智能计算能力评测指标、智能任务评测指标、可靠性评测指标等。

1 智能软件的可靠性

智能软件可靠性是指系统在既定运行条件与限定时间范围内,稳定、准确实现预设功能,且不出故障、异常及失效问题的综合性能^[6]。从影响维度划分,智能软件可靠性主要涵盖模型、数据、运行平台三大核心层面,各层面独立作用且共同决定智能软件的整体运行稳定性。其中,模型层面聚焦算法与代码的自身合理性,数据层面贯穿训练数据与实时环境数据的质量特性,平台层面则指向软硬件运行载体的支撑稳定性。

(1) 模型可靠性

模型可靠性是针对智能软件内嵌机器学习、深度学习模型的核心可靠性评价指标,用于判定模型算法运行与决策输出的稳定合规能力^[7],具体包含五项核心特性:

功能正确性体现模型对既定业务任务的适配与执行能力,衡量模型任务输出结果的精准度;可解释性指代模型针对各类业务场景作出决策的逻辑可信度,反映决策过程的可追溯、可研判程度;鲁棒性是模型应对异常输入、突发场景的抗干扰能力,可有效规避非常规工况下的模型失效问题;容错性体现模型在局部故障、参数偏差等问题存在时,维持标准性能输出、满足使用需求的能力;易恢复性则表征模型发生失效故障后,能够在短时间内自主修复、回归正常运行状态的能力。

(2) 数据可靠性

数据是智能模型训练与迭代的核⼼基础,数据可靠性直接决定模型训练效果与落地性能^[8],其核心评价标准围绕数据集质量展开,主要包含三项关键指标:

数据均衡性用于评判数据集各类别样本的分布均匀程度,避免因样本倾斜导致模型训练偏差;数据规模性指参与模型训练的样本总量,充足的数据体量是保障模型泛化能力的前提;数据标注质量衡量数据集样本标签的精准度与规范性,高质量的标注数据能够从源头规避模型学习偏差,保障模型训练的有效性。同时,实际应用中的对抗性数据、分布偏移数据、野值异常数据,也会直接影响数据整体可靠性,进而干扰模型实际运行效果。

(3) 平台可靠性

平台可靠性是保障智能软件稳定运行的底层支撑条件,主要衡量软件运行所依托的全套软硬件载体的稳定性能^[9],分为软件平台与硬件平台两大维度:

软件平台可靠性聚焦智能软件运行依赖的软件生态,涵盖操作系统、深度学习框架、大数据运行平台等基础软件工具的稳定性与适配性;硬件平台可靠性则针对物理运行载体,包括 AI 芯片、各类传感采集设备等硬件设施的运行精度、稳定性与抗损能力,软硬件平台的协同稳定是智能软件持续可靠运行的基础保障。

2 智能软件可靠性综合评测指标体系

为了能够充分评测计算机的计算能力及智能软件的可靠性,本文针对目标识别、目标检测、智能决策等典型应用,构建了一套智能软件可靠性综合评测指标体系模型。如图1所示,该指标体系模型包括通用性能评测指标、智能计算能力评测指标、智能任务评测指标、可靠性评估指标等四个方面。

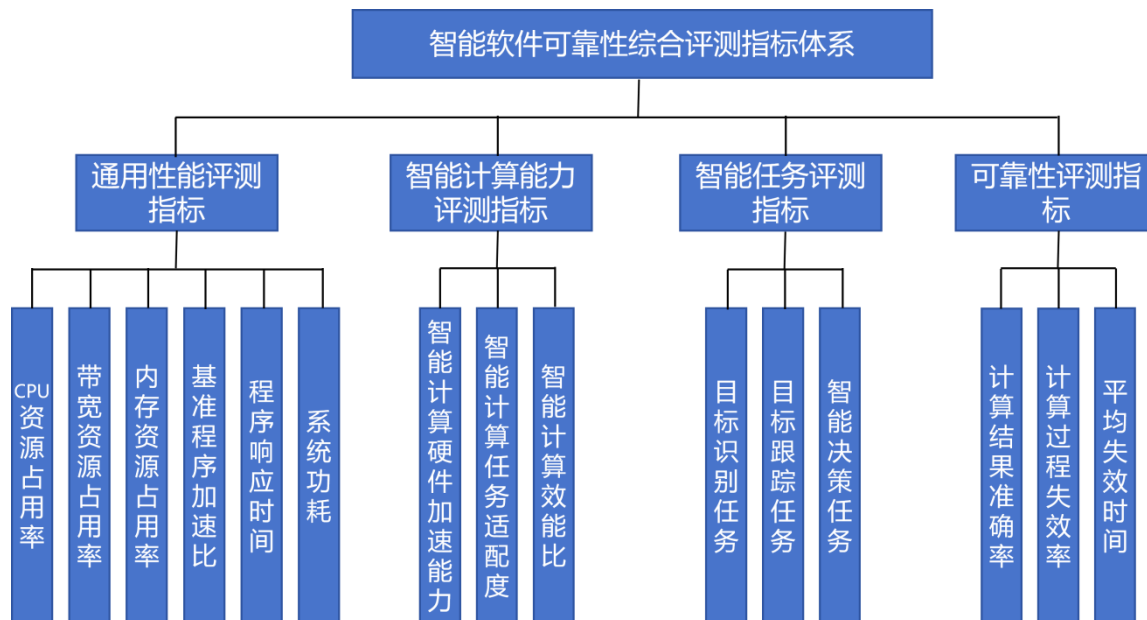


图1 智能软件可靠性评测指标体系

Figure 1 Index System for Reliability Evaluation of Intelligent Software

2.1 通用性能评测指标

通用性能评测指标用于评估智能计算平台的通用性能,包括以下六个具体指标:

(1) CPU 资源占用率

CPU 资源占用率反映了运行的智能软件所占用的 CPU 资源,表示机器在某个时间点运行程序的情况。通过 CPU 资源占用率可以看出系统在运行时的 CPU 负载情况,通过计算机系统运行时的负载来判断该智能计算平台的 CPU 性能以及针对特殊任务的优化情况。

(2) 带宽资源占用率

带宽资源占用率反映了智能计算平台在运行软件时的带宽占用情况。通过带宽资源占用率可以计算出系统的带宽负载情况,进而判断该智能计算平台的网络带宽性能以及针对目标任务的网络优化。带宽资源占用率高,则表明网络传输很繁忙。如果智能系统的响应时间越来越大,而相关程序带宽资源占用率始终下不去,说明网络传输遇到了瓶颈,系统需要调优。

(3) 内存资源占用率。

内存资源占用率是软件测试领域一种常用的性能评测指标,反映了运行的软件占用内存资源的情况。通过内存资源占用率的指标,可以看出系统在运行时的内存负载情况,并判断如何对系统资源进行优化与调度。

(4) 基准程序加速比

通过基准程序加速比可以清楚地了解到待测智能计算机相对于基准智能计算机系统在运行基准程序上的性能优化情况。在计算基准程序加速比时,先要通过代码获取基准程序的执行时间 T_{mark} ,再通过代码插桩的方式得到待测程序的执行时间 T_{target} ,则基准程序加速比可通过下式计算得到。

$$Radio = 1 - \frac{T_{target}}{T_{mark}} \quad (1)$$

(5) 程序响应时间

程序响应时间是指从用户发出指令到待测试程序开始执行所花费的时间，通过程序响应时间指标可以看到该智能计算机在任务调度方面的优化情况。程序的响应时间可以通过一个代理程序调用待测程序的来获取。

(6) 系统功耗

系统功耗反映了智能计算机在运行待测试程序时的功耗情况，通过该指标可以看出智能计算平台在硬件和软件方面对功耗的优化情况。智能计算平台的功耗获取需要嵌入式硬件的支持，可以通过功耗计来获取智能计算平台在运行待测试程序时的功耗情况。

2.2 智能计算能力评测指标

智能计算能力评测指标用于评估智能计算平台在运行智能软件时的智能计算能力，主要包括以下三个指标：

(1) 智能计算硬件加速能力

智能计算硬件加速能力指的是在实际智能计算任务中，智能计算平台相较于基准计算机的加速能力。在计算智能计算硬件加速比时，先要通过代码获取基准智能任务的执行时间。

(2) 智能计算任务适配度

智能计算任务适配度指的是智能计算平台对于目标识别任务、目标检测任务、决策任务的加速能力，其反映了该智能计算平台针对不同智能任务的优化适配情况。

(3) 智能计算效能比

智能计算效能比用于对智能任务计算耗能和智能任务执行效率进行综合评估。该指标可以反映出智能计算平台针对特定智能任务的优化情况。智能计算效能比会分别对目标识别任务、目标检测任务、决策任务进行评估。

2.3 智能任务评测指标

智能任务评测指标用于评估智能模型的输出结果，本文设计的智能软件可靠性评测指标体系主要面向目标识别、目标检测、智能决策等三种任务。

(1) 目标识别任务评测指标包括准确率、精确率、召回率、F1 调和值、置信度、AUC 值、检测时间等，用于评测智能模型在不同情况下执行目标识别任务的性能。

(2) 目标检测任务评测指标包括检测精度、重合度、准确率、检测时间等，主要用于评测智能模型在不同情况下执行目标检测任务的性能。

(3) 智能决策任务评测指标包括响应时间、迭代次数、完成度等指标，主要用于评测智能模型在做任务决策时的性能。

2.4 可靠性评测指标

可靠性评测指标主要由计算结果准确率、计算过程失效率、平均失效时间这三部分构成。

(1) 计算结果准确率

计算结果准确率指标可以反映智能计算平台在对软硬件进行优化时，是否存在会导致智能任务准确率下降的优化方法。为了获取计算结果准确率，首先需要给所有的基准测试程序设置测试预言 (Test Oracle)。测试预言是检验待测系统在特定执行下是否正确运行的方法，是软件测试过程中的

重要环节。在获取计算结果准确率时,测试预言可以自动化地获取程序的执行情况。

(2) 计算过程失效率

软件失效是指软件运行时产生的一种不希望或不可接受的外部行为结果。计算过程失效率可以有效的反映智能软件在运行过程中出现的失效次数及比例。

(3) 平均失效时间

平均失效时间用于统计每次失效的平均时间,同计算过程失效率一样,平均失效时间也能反映智能软件的失效情况。

3 结论

智能软件以深度神经网络为核心,具备数据驱动的特征,其可靠性不再单纯取决于代码逻辑,而是由模型、数据、软硬件运行平台三者共同决定,传统软件可靠性测试手段已经难以满足智能软件质量评价需求。本文剖析了智能软件可靠性的三层核心影响因素,建立一套面向目标识别、目标检测、智能决策场景的智能软件可靠性综合评测指标体系,涵盖通用性能、智能计算能力、智能任务、可靠性评估四大类指标,覆盖平台资源开销、硬件加速效能、业务任务输出精度、运行失效统计等多项具体度量项。

该指标体系能够实现对智能软件从底层运行平台到上层业务任务的可靠性评测,可以为智能软件的测试验证、版本迭代和质量管控提供可落地的参考框架。但本文仅完成指标体系模型的理论构建,后续研究将基于该指标体系设计量化评分方法,搭建对应的评测实验环境,通过多组智能软件实例开展实证测试,进一步完善指标权重,验证该评测指标体系的实用性与有效性。

参考文献

- [1] 高萍萍.大数据驱动下的智能软件开发框架研究[J].信息与电脑,2026,38(9):207-209.
- [2] He W, Li B, Song D. Decision Boundary Analysis of Adversarial Examples[C]. International Conference on Learning Representations. 2018,1-15.
- [3] Guo J., Bao W., Wang J., et al. A Comprehensive Evaluation Framework for Deep Model Robustness[J]. Pattern Recognition, 2023, 137: 109308.
- [4] Zhang D., Dong Y. Adv-BDPM: Adversarial Attack Based on Boundary Diffusion Probability Model[J]. Neural Networks, 2023, 167: 730-740.
- [5] Ukita J, Ohki K. Adversarial Attacks and Defenses Using Feature-Space Stochasticity[J]. Neural Networks, 2023.
- [6] Pons L., Ozkaya I. Priority Quality Attributes for Engineering AI-enabled Systems[J]. arXiv.1911.02912, 2019.
- [7] Goldstein B F, Srinivasan S, Das D, et al. Reliability evaluation of compressed deep learning models[C]//2020 IEEE 11th Latin American Symposium on Circuits & Systems (LASCAS). IEEE, 2020: 1-5.
- [8] Kumar R, Kumar S, Tiwari S K. A study of software reliability on big data open source software[J]. International Journal of System Assurance Engineering and Management, 2019, 10(2): 242-250.
- [9] Alshemali B, Kalita J. Improving the reliability of deep neural networks in NLP: A review[J]. Knowledge-Based Systems, 2020, 191: 105210.

基金项目: 南通市自然科学基金(JC2023023)。

^{1,*}第1作者&通讯作者简介: 高德鹏(1988-),男,博士,讲师,研究方向:人工智能、软件工程、计算机视觉等。E-mail: gaodepeng@ntit.edu.cn。