

Quality Assurance of AI-Enabled Cyber-Physical Systems

呂, 徳運

<https://hdl.handle.net/2324/7363810>

出版情報 : Kyushu University, 2024, 博士 (情報科学) , 課程博士
バージョン :
権利関係 :



氏 名 : 呂 徳運 (LYU DEYUN)

論 文 名 : Quality Assurance of AI-Enabled Cyber-Physical Systems
(AI 対応のサイバーフィジカルシステムの品質保証)

区 分 : 甲

論 文 内 容 の 要 旨

Cyber-Physical Systems (CPSs) have been broadly deployed in safety-critical domains, such as autonomous driving, medical devices, etc. In recent years, Artificial Intelligence (AI), mainly deep neural networks (DNN), has been increasingly adopted to control CPSs, giving birth to AI-enabled CPSs. Despite the great potential in terms of control, the safety of AI-enabled CPSs becomes a major concern. Due to the inherent uncertainty and opacity nature of DNNs, a DNN controller may produce unforeseen and inexplicable behaviours that lead the plant to malfunctions and catastrophic hazards. Hence, it is essential to detect the potential unsafe behaviours of DNN controllers and identify the root causes that produce them. To enhance the reliability of AI-enabled CPSs, fixing the unsafe system behaviours is also necessary after identifying DNN parameters responsible for the violation, which has not received much attention. Besides, CPSs are often kept as private intellectual properties by industrial practitioners, very few benchmarks of AI-enabled CPSs are publicly available for quality assurance research.

To address these challenges, this dissertation explores the entire lifecycle of quality assurance of AI-enabled CPSs (i.e., benchmarking, falsification, fault localisation and repair), and makes four key contributions:

- Benchmarking AI-enabled CPSs. We present a public benchmark of industry-level CPS in seven domains and build AI controllers for them via state-of-the-art deep reinforcement learning (DRL) methods. Based on that, we further perform a systematic evaluation of these AI-enabled systems with their traditional counterparts to identify current challenges and future opportunities.
- Coverage-guided falsification framework specifically designed for AI-enabled CPSs. We proposed a coverage-guided falsification framework FalsifAI, which exhaustively employs eight time-aware coverage criteria as guidance to explore the temporal behaviours of AI-enabled CPS. These coverage criteria aim to capture different time-series features of DNN behaviours and provide better guidance to find violation cases to the system specification. The large-scale evaluation demonstrates the effectiveness of FalsifAI and eight time-aware coverage criteria.
- Fault localisation of AI-enabled CPSs. We propose SpectAcle, that can localise the faulty parameters in DNN controllers. SpectAcle considers the DNN inferences preceding the specification violation and uses forward impact to determine the DNN parameters that are more relevant to the DNN outputs. Then, it identifies which of these parameters are responsible for the specification violation, by adapting classic suspiciousness metrics. Moreover, we propose two versions of SpectAcle, that consider differently the timestamps that precede the specification violation.
- Search-based repair of DNN controllers. We propose a novel DNN repair approach guided by system-level specifications. The approach takes as input a system-level specification, some tests violating the specification, and some faulty DNN weights. The approach searches for alternative weight values with the goal of fixing the behaviour on the failing tests without breaking the passing tests. We also propose a heuristic that allows us to accelerate the search by avoiding the execution of some tests.

To summarize, this dissertation spans the entire lifecycle of building reliable AI-enabled CPSs, advances the understanding and deployment of AI in CPS, paving the way for more trustworthy AI-enabled CPSs.