March 17, 2023

Techniques for Effectively and Efficiently Mitigating Risks by Misperception of Image Recognition AI

-- Evaluated in Safety Benchmark for Automated Driving Systems --

The research team led by ISHIKAWA Fuyuki at the National Institute of Informatics (NII, Japan) developed techniques for effectively and efficiently mitigating risks by misperception of image recognition AI with the research team led by MA Lei at Kyushu University. This research was conducted under the MIRAI-eAI project (*1) funded by the Japan Science and Technology Agency (JST, Japan).

A critical issue of Deep Neural Network (DNN) techniques is unexpected performance regression. A fix to improve a certain type of misperceptions may lead to regression in other types as an enormous number of parameters intricately affect the perception of different object types.

In the eAI project, we have investigated how to repair image recognition DNN in a controlled way by combining DNN repair techniques of different roles. Specifically, we developed techniques to explore causes and fixes for different misperception types (NII), to suppress unintended regressions by analyzing the history of the target DNN (NII), and to update the structure (architecture) of DNN not only parameter values (Kyushu Univ.).

In our experiments for automated driving systems, we evaluated the techniques with safety benchmarks defined by safety expects, especially, those from automotive companies. We demonstrated effective and efficient risk mitigation by controlled DNN repair with preserving many safety requirements.

We plan to integrate the repair techniques as a framework and work on industrial demonstration based on different demands in individual companies such as visions and policies for automated driving as well as characteristics of driving data.

The research outcomes from NII and Kyushu Univ. are presented in ICST 2023(*2), SANER 2023(*3), and TOSEM(*4), flagship conferences and journals in software engineering.

Background

Image recognition AI plays significant role in automated driving systems or advanced driver assistant sytems for the detection and classification of objects such as pedestrians we well as road signs and lanes. For implementing image recognition AI, deep learning techniques are used, in which millions or more of parameters in a computation model called deep neural network (DNN) are automatically configured through training on a dataset.

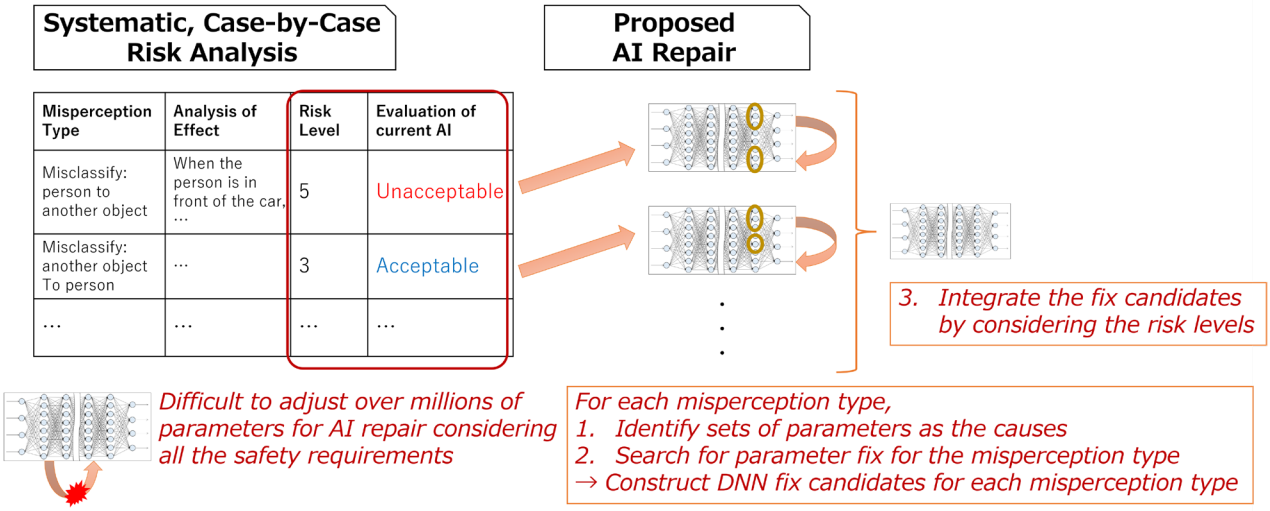
Research Organization of Information and Systems National Institute of Informatics
Kyushu University
Japan Science and Technology Agency (JST)

On the other hand, it is essential to certify that risks due to various errors and faults are acceptably small in safety-critical systems such as automated driving systems. The design and configuration of the target system are explored to mitigate the risks by systematically and individually analyzing how severe hazards each type of error or fault leads to, and in what situations. It is thus necessary to assess and mitigate the risks due to misperceptions of AI for different objects and environments. This principle has been widely requested as it is shown in recent guidelines for the quality of AI systems (AIQM(*5), QA4AI(*6)) as well as standards for safety of autonomous products (ISO 21448(*7), ANSI/UL 4600(*8)).

However, there have been large difficulties in fixing misperception with existing DNN techniques. Suppose engineers found the trained DNN model has a high misperception rate for "misclassifying a pedestrian to another object" and intend to fix it. Generally, the common approach is to collect additional training data for the misperception type and have retraining of the DNN model. However, the outcome of retraining is often unexpected as millions of parameter values are "shuffled" in the DNN model. This point requires intensive effort of trial and error to obtain the expected fix. In addition, a fix for one misperception type may cause regressions in other types, i.e., the updated model may have misperceptions of inputs for which it succeeded. In this way, existing DNN techniques have difficulties in controllability of update and repair for mitigating multiple, specific types of misperceptions.

Research at NII

To address the problem, the NII research team investigated a DNN repair technique that explores sets of parameters responsible for each misperception type. This technique first identifies sets of parameters that are causes of each misperception type such as "misclassifying a person to rider" and "misclassifying train to bus" by using a technique called fault localization. Then, parameter fixes are explored to fix each misperception type. Finally, the obtained candidate fixes for different misperception types are merged by considering the risk levels of the misperception types (Figure 1), thus achieving more effective and efficient risk mitigation than existing techniques.



<Figure 1> Integration Mechanism with Parameter Fix Search and Risk Levels

The advantage of this technique was evaluated by a prototype of a safety benchmark defined in a working group involving practitioners from automotive companies and other companies. This benchmark defines the holistic risk score by classifying 12 types of misperception into 3 risk levels. The evaluation result demonstrated the feasibility of DNN repair on multiple misperception types, which was difficult with existing techniques. This result will be presented in ICST 2023(*2), a flagship conference of software testing, in April 2023.

The NII research team also worked together with a team in Fujitsu Limited on the research of a DNN repair technique that prevents unintended regressions of prediction performance for significant target objects (presented in March 2022 (*9)). This technique makes use of insights obtained from analysis of past versions of DNN to suppress regressions in prediction performance for significant objects. In Autumn 2022, the effectiveness of regression-controlled repair techniques was empirically confirmed in experiments with a social security AI application. These experimental results will be presented in the industry track of SANER 2023(*3), a flagship conference of software analysis, in March 2023.

Research at Kyushu Univ.

The team at Kyushu Univ. has investigated different approaches to the DNN repair techniques, led by Assoc. Prof. MA Lei and Prof. ZHAO Jianjun. In the technique presented in 2021 (*10), a technique called styler transfer was used to attach features of failed input images to training data for handling unknown noise patterns detected in operation. This technique thus realized DNN repair that handles failure patterns that cannot be explicitly stated by engineers and unexpected noise distributions in operation. In another study of 2022, more potential of DNN

repair is explored by modifying the structure (architecture) of DNN not only parameter values. This approach can be combined with other repair techniques such as the ones by NII and Fujitsu. This research on architecture-level DNN repair will be published in ACM TOSEM [(*4)], a flagship journal in software engineering, in 2023.

Achievement and Perspective in eAI Project

The eAI project, involving the research teams of NII, Kyushu Univ. and Fujitsu, has promoted the development of "AI repair tools" given different use cases and demands in the industry. The technical development and evaluation have been driven by industrial demands in the form of iterated safety benchmarks defined by safety experts.

We will conduct full-fledged experiments towards image recognition AI that complies with fine-grained, multiple safety requirements. This effort aims at meeting various demands for quality, safety, and reliability in automated driving by integrating multiple AI repair techniques of different roles. Specifically, we will work on industrial demonstrations based on different demands in individual companies such as visions and policies for automated driving as well as characteristics of driving data.

The eAI project also involves another axis of techniques for reliable AI constriction with small data and its experiments in the healthcare domain, led by Prof. SUZUKI Kenji at Tokyo Institute of Technology. In addition, the project also involves research and industrial experiments on a framework for the development and operation of AI systems, led by Prof. WASHIZAKI Hironori at Waseda University. The framework supports the end-to-end process cycles from fine-grained requirements and risk analysis, construction/repair of AI systems, and to the operation. Some of the results were presented in IEEE Computer, the flagship magazine of computer science [(*11)]. Through these activities, the eAI project aims at establishing engineering methodologies for AI systems that can be tailored to meet the fine-grained demands of the industry and society.

Comment by ISHIKAWA Fuyuki, Project Leader of eAI project

"We started the eAI project given the strong demand for quality and engineering techniques for AI systems based on machine learning, especially, deep learning techniques. The research teams of NII, Kyushu Univ., and Fujitsu have focused on the significant problem of AI repair, which comes after we obtain AI systems that "works some well." In the engineering of traditional software systems, repair or "debugging", including the goal of suppressing regressions, has been a most error-prone and costly task. It is inevitable to tackle similar

difficulties of AI repair in safety-critical systems in which fine-grained risk analysis and assessment are essential.

NII has conducted intensive research on the safety of driving behavior of automated driving systems in the ERATO-MMSD project led by HASUO Ichiro [(*12)]. The eAI project complements and collaborates with it by focusing on the perception AI with high uncertainty. We will deepen the discussion and experiments with automotive companies and enhance the DNN repair techniques for increased safety. We will also work on surrounding techniques combined with the perception AI for a comprehensive framework for certification and improvement of safety for automated driving systems."

Funding

The presented research is supported by the project for "Engineerable AI Techniques for Practical Applications of High-Quality Machine Learning-based Systems" (JPMJMI20B8), in the "Super Smart Society (Society 5.0)" mission area, JST-Mirai Program.

Title and Authors (NII)

Title : Distributed Repair of Deep Neural Networks
Authors : Davide Li Calsi, Matias Duran, Xiao-Yi Zhang, Paolo Arcaini, Fuyuki Ishikawa
Venue : The 16th IEEE International Conference on Software Testing, Verification and Validation (ICST 2023)
Date : April 16-20, 2023 (in Ireland)

Title and Authors (Kyushu Univ.)

Title : ArchRepair: Block-Level Architecture-Oriented Repairing for Deep Neural Networks
Authors : Hua Qi, Zhijie Wang, Qing Guo, Jianlang Chen, Felix Juefei-Xu, Fuyuan Zhang, Lei Ma, Jianjun Zhao
Journal : ACM Transactions on Software Engineering and Methodology (TOSEM)
Date : 2023 (to appear)

Title and Authors (NII and Fujitsu)

Title : An Experience Report on Regression-Free Repair of Deep Neural Network Model
Authors : Takao Nakagawa, Susumu Tokumoto, Shogo Tokui, Fuyuki Ishikawa
Venue : The 30th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2023, Industry Track)
Date : March 23, 2023 (Thursday) (in Macao)

<Media Contact>

Research Organization of Information and Systems

National Institute for Informatics

Publicity Team

TEL : +81 (0)3-4212-2164

E-mail : media@nii.ac.jp

Kyushu University

Public Relations Office

TEL : +81 (0)92-802-2130

E-mail : koho@jimu.kyushu-u.ac.jp

〈About JST Projects〉

Japan Science and Technology Agency (JST)

Department of R&D for Future Creation

KOIZUMI Terutake

TEL : +81 (0)3- 6272-4004

E-mail : kaikaku_mirai@jst.go.jp

(*1) MIRAI eAI Project: a research project funded by the "Super Smart Society (Society 5.0)" mission area, JST-Mirai Program. It tackles techniques for the construction and repair of AI systems to meet fine-grained requirements for safety and reliability. Its goal is proof-of-concept in the two areas of healthcare and transportation. The official name is “Engineerable AI Techniques for Practical Applications of High-Quality Machine Learning-based Systems”, abbreviated as eAI project. https://engineerable.ai/

(*2) ICST 2023 : The 16th IEEE International Conference on Software Testing, Verification and Validation. The “A” rank in the conference ranking called CORE.

(*3) SANER 2023 : The 30th IEEE International Conference on Software Analysis, Evolution and Reengineering. The “A” rank in the conference ranking called CORE.

(*4) TOSEM : ACM Transactions on Software Engineering and Methodology. The “A*” rank in the journal ranking called CORE.

(*5) AIQM : Machine Learning Quality Management Guideline. A guideline for the quality of AI systems organized by The National Institute of Advanced Industrial Science and Technology (AIST). Normative statements are organized regarding quality characteristics and levels for AI systems. https://www.digiarc.aist.go.jp/en/publication/aiqm/

(*6) QA4AI : Guideline for Quality Assurance of AI-based products and services. A guideline organized by experts of software quality and testing techniques, involving concrete activities and planning from the viewpoint of engineers. https://www.qa4ai.jp/

(*7) ISO 21448 : A standard that focuses on the safety of autonomous functions that make use of sensors and actuators, not safety regarding systematic faults, called SOTIF (safety of the intended functionality).

(*8) ANSI/UL 4600 : A standard that states a framework for safety argument (called safety case) of autonomous products such as ones that use machine learning techniques.

(*9) Tokui et al., NeuRecover: Regression-Controlled Repair of Deep Neural Networks with Training History, The 29th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2022)

(*10) Yu et al., DeepRepair: Style-Guided Repairing for Deep Neural Networks in the Real-World Operational Environment, IEEE Transactions on Reliability, Vol. 71 No. 4, 2021

(*11) Washizaki et al., Software-Engineering Design Patterns for Machine Learning Applications, IEEE Computer, Vol. 55 No. 3, 2022

(*12) ERATO Hasuo Metamathematics for Systems Design (ERATO-MMSD): a project funded by the Exploratory Research for Advanced Technology (ERATO) scheme of the Japan Science and Technology Agency (JST). The project conducts research for the quality assurance of cyber-physical systems, specifically automated driving systems, with reliability techniques for their modeling, formal verification, and testing.
https://www.jst.go.jp/erato/hasuo/en/
The past NII press release includes the following:
- Automated Technique to Efficiently Discover Severe Problems in Automated Driving Systems
https://www.nii.ac.jp/en/news/release/2021/1115.html
- A New Method for Mathematically Proving the Safety of Automated Driving Vehicles
https://www.nii.ac.jp/en/news/release/2022/0707.html