International Journal of Advanced Smart Convergence Vol.10 No.3 59-65 (2021)
http://dx.doi.org/10.7236/IJASC.2021.10.3.59

IJASC 21-3-8

Roadmap Toward Certificate Program for Trustworthy Artificial Intelligence

Min-gyu Han[1] and Dae-Ki Kang[2]

[1]Professor, ICT Convergence Program, Hansung University, Seoul, South Korea
[2]Professor, Dept. of Computer Engineering, Dongseo University, Busan, South Korea
[1]andyhan@hansung.ac.kr, [2]dkkang@dongseo.ac.kr

Abstract

In this paper, we propose the AI certification standardization activities for systematic research and planning for the standardization of trustworthy artificial intelligence (AI). The activities will be in two-fold. In the stage 1, we investigate the scope and possibility of standardization through AI reliability technology research targeting international standards organizations. And we establish the AI reliability technology standard and AI reliability verification for the feasibility of the AI reliability technology/certification standards. In the stage 2, based on the standard technical specifications established in the previous stage, we establish AI reliability certification program for verification of products, systems and services. Along with the establishment of the AI reliability certification system, a global InterOp (Interoperability test) event, an AI reliability certification international standard meetings and seminars are to be held for the spread of AI reliability certification. Finally, TAIPP (Trustworthy AI Partnership Project) will be established through the participation of relevant standards organizations and industries to overall maintain and develop standards and certification programs to ensure the governance of AI reliability certification standards.

Keywords: *Trustworthy Artificial Intelligence, AI Reliability, AI Certification, IoT platform convergence*

1. Introduction

Artificial intelligence (AI) technology has become a key technology to dominate the 4th industrial revolution and most contemporary technologies are getting more and more dependent on AI. Thus, there has been increase about the concern on malfunctioning of AI. More specifically, there are significant problems in AI learning algorithms and AI implementation mechanisms in terms of their robustness. For example, adversarial examples can easily deceive AI systems with adversarial perturbation.

Regarding the malfunctioning, problems from misclassification due to adversarial examples are:

- Inducing decision error of service systems using AI. This can be protected by deceptive attack and defense.
- Unauthorized access to personal identification and authentication system. This can be addressed

Manuscript Received: July. 9, 2021 / Revised: July. 13, 2021 / Accepted: July. 15, 2021
Corresponding Author: dkkang@dongseo.ac.kr
Tel: +82-51-320-1724, Fax: +82-51-327-8955
Professor, Department of Computer Engineering, Dongseo University, South Korea

by detecting abnormal access.

- Inability to protect AI engines from illegal and harmful data. This should be fundamentally treated by protecting federated attack/defense.

As will be discussed in the related work, there have been numerous academic research on trustworthy artificial intelligence technology, however, there have been very little systematic research and planning for the standardization of trustworthy artificial intelligence (AI). Moreover, constructing reliable certification program for AI engines will be beneficial to the industrial area in terms of standardization and economics viewpoints. Against this background, in this paper, we propose a systematic and detailed plan for building certification program of trustworthy artificial intelligence, with which we hope to assist Korea national policy.

2. Related work

There have been many interesting research on deceptive attacks and protection techniques. Guo et al. presented a simple method for the construction of adversarial images in the black-box setting [1]. Their method used a principle that they could randomly sample a vector from a predefined orthonormal basis. After that they could either add/subtract it to the target image. Lee et al. proposed a simple and effective algorithm for detecting abnormal samples [2]. Their algorithm could be applicable to any pre-trained softmax neural classifier. They obtained the class conditional Gaussian distributions with respect to feature maps of the neural network models under Gaussian discriminant analysis. The distributions produced a confidence score based on Mahalanobis distance. Huang and Zhang presented a novel method for black-box adversarial attack, which learned a low-dimensional embedding from pretrained models, and then performed efficient search within the embedding space to attack an unknown target network [3].

For detecting abnormal access, Kwon et al. presented a selective poisoning attack that reduces the accuracy of only chosen class in the model [4]. Their proposed method reduced the accuracy of a chosen class in the model by training malicious training data corresponding to a chosen class, while maintaining the accuracy of the remaining classes. Raghunathan proposed a method based on a semidefinite relaxation that outputs a certificate for a given network and test input [5]. The certificate ensured that no attack could force the error to exceed a certain value. Because the certificate was differentiable, they jointly optimized it with the network parameters, providing an adaptive regularizer that encouraged robustness against all attacks. Shafahi et al. presented an optimization-based method for crafting poisons [6]. They showed that one single poison image could control classifier behavior when transfer learning is used.

For attacking and defending federated learning, Bagdasaryan et al. proposed a new model-poisoning methodology based on model replacement [7]. Yin et al. developed distributed learning algorithms that are provably robust against Byzantine failures [8]. Zhu et al. showed that it was possible to obtain the private training data from the publicly shared gradients [9]. They called that leakage as Deep Leakage from Gradient and empirically validated the effectiveness on both computer vision and natural language processing tasks.

3. Standardization of trustworthy artificial intelligence

3.1 Topics for trustworthy artificial intelligence

We envision that there are five issues for trustworthy artificial intelligence.

1. First of all, we work toward complete report on classifications of the attack methods including the novel attack techniques. Those attack methods include attacks on object segmentation, image classification, and deceptive speech/text attacks. Specific task plans for trustworthy artificial intelligence will be one of our further research directions.

2. Secondly, we formalize AI model framework and machine learning models for the purpose of standardization. After that, we will analyze their vulnerabilities and use cases. Finally, we will establish the requirement specifications for robustness of the framework and models.
3. We detect dysfunction security vulnerabilities of AI models and to define structures of defense techniques. Finally, we establish functional specifications for them.
4. We define four grades of stability evaluation tools to find use cases and to diagnose vulnerabilities of AI systems. This task is similar to grading system of Cryptographic Module Validation Program (CMVP), developed by NIST (National Institute of Standards and Technology). The purpose of this task is the establishment of test technical standards and certification standards.
5. We apply vertical domain of AI reliability technology standards. These vertical domains include IoT technology and Contents-related technology.

3.2 Task plan for trustworthy artificial intelligence

With those vision items as a long-term goal, we believe the following actions should be performed for trustworthy AI standardization.

1. Response activities of the policy committee at the governance level of the international organizations for standardization. These international organization activities include ISO/IEC JTC1, oneM2M TP response activities. (Note that EIC stands for International Electrotechnical Commission, JTC1 is Joint Technical Committee 1, and TP is Technical Plenary.)
2. Establishment of committee/group of international standardization organization and production of chairperson. (Note that WI is Work Items, and WG is Working Groups.)
   A. Elected as WI rapporteurs in Trustworthy AI group of ISO/IEC JTC1 WG3
   B. Establishment of Artificial Intelligence IoT WG in oneM2M and elected as a chairperson
   C. Establishment of Trustworthy AI certification system for AI reliability-based technology standardization and certification program operation
3. Establishing a proactive standard concept for future new technologies. This includes establishment of AI engine reliability evaluation program and test certification technology standard.
4. Discovering the future ICT standardization agenda that Korea can lead. (ICT stands for Information and Communications Technology.)
   A. Establishment and governance of AI engine reliability evaluation certification program
   B. Expanding to multiple vertical industries
5. Establishing a global standard cooperation network for international cooperation in the standard field and domestic invitation of international conferences
   A. Establishment of Trustworthy AI Partnership Project/Program (TAIPP) system of regional standardization organizations
   B. Domestic standard meeting invitation: ISO/IEC JTC1, oneM2M
6. Proposal and approval of standards through technical committee activities
   A. Standards for AI deceptive attack detection and defense
   B. AI reliability assessment methods and criteria standards

3.3 Target standards bodies

The first target standardization body for our plan is “ISO/IEC JTC1 SC42 Artificial Intelligence” group. (SC stands for Sub-Committee.) The group includes two working groups regarding this research. The first one is WG3 for trustworthiness which is interested in development of standards related to reliability of artificial

intelligence (reliability overview, robustness, bias, ethics, risk management, etc.). The second one is WG4 about use cases and applications, which is interested in identifying use cases of artificial intelligence.

The second target standardization body for our plan is “oneM2M Artificial Intelligent IoT (A-IoT)”. European Telecommunications Standards Institute (ETSI) and oneM2M have conducted basic research for AI convergence of IoT. The technical reports of this effort include oneM2M WI-0105 for system enhancements to support AI capabilities, ETSI TR 103 674 for artificial intelligence and the oneM2M architecture (included in SmartM2M), and ETSI TR 103 675 for AI for IoT: A Proof of Concept (also included in SmartM2M). Their AI/ML use-cases provide point-solution and horizontal capabilities. Specifically, the use case #8 of them is for “Trustworthy AI”, and the use case #9 is for “Verifiable AI”, which are related with the proposed research.

4. AI certification standard activities

We present a general flow chart of AI certification standard activities in Figure 1.

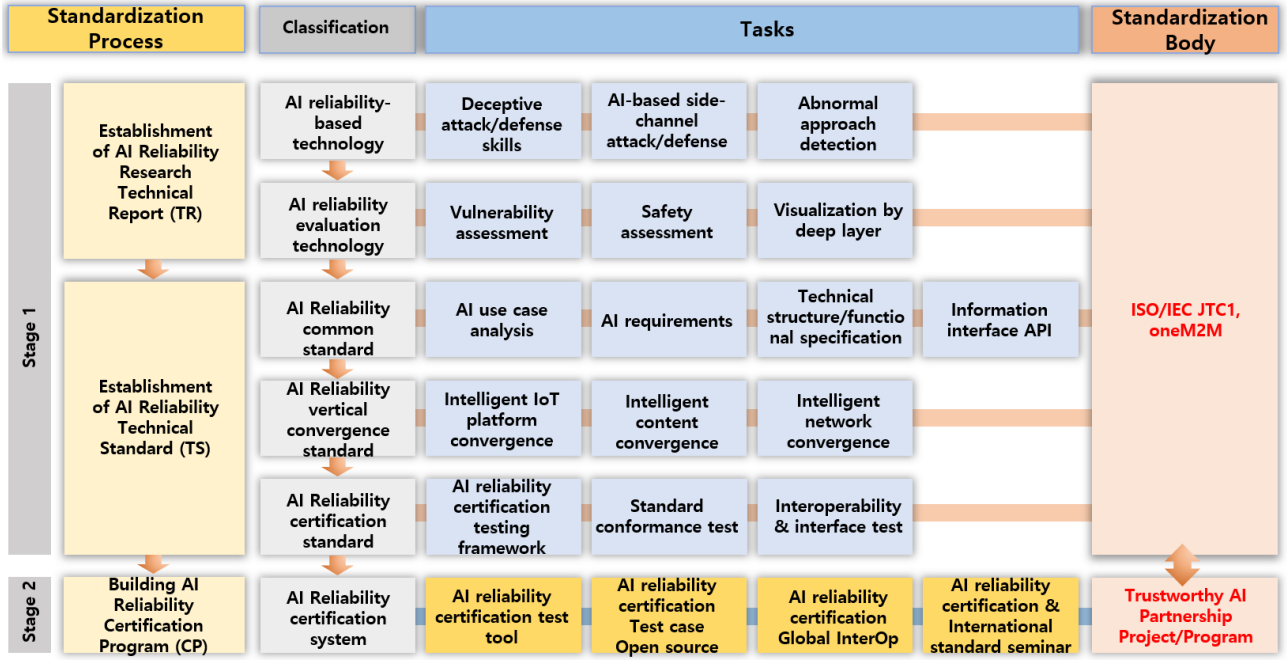

Figure 1. General flow of AI certification standard activities

For stage 1, we establish AI reliability research technical report and AI reliability technical standard. We categorize AI reliability research into AI reliability-based technology and AI reliability evaluation technology.

When we perform research on AI reliability-based technology, we will focus our research on deceptive attack/defense skills, AI-based side-channel attack/defense, and abnormal approach detection. The research area of AI reliability evaluation technology will be further divided into vulnerability assessment, safety assessment, and layer visualization.

The establishment of AI reliability evaluation technology will be followed by AI reliability technology standards, which will be further divided into AI reliability common standard, AI reliability vertical convergence standard, and AI reliability certification standard. Those three standards will be exercised sequentially. For AI reliability common standard, we will investigate AI use case analysis, AI requirements, technical structure/functional specification, and information interface API. AI reliability vertical convergence standard consists of intelligent IoT platform convergence, intelligent content convergence, and intelligent

network convergence. Finally, AI reliability certification standard is focused on AI reliability certification testing framework, standard conformance test, and interoperability & interface test.

Again, note that these standardization activities will be performed in ISO/IEC JTC1, and oneM2M standardization organizations.

After stage 1, we plan to build AI reliability certification program (CP) in stage 2. The program will include AI reliability certification test tool, AI reliability certification regarding test case and open source, AI reliability certification regarding global interop, and AI reliability certification & international standard seminar. These activities will be administrated by our Trustworthy AI Partnership Project/Program.

4.1 Stage 1: Establishment of AI reliability certification standard

We detail the action items for the stage 1 as follows:

1. A study on the scope and possibility of standardization of AI reliability-based technology
   A. Base attack/defense technology to ensure the reliability of the AI engine
   B. AI-based side-channel attack/defense, abnormal approach detection technology
2. A Study on the standardization scope and possibility of AI reliability evaluation technology
   A. AI vulnerability and stability evaluation technology and in-depth layer-by-layer monitoring technology through visualization
3. Activities to establish standards for AI reliability common technology
   A. Requirement derivation through AI use case analysis, technical structure design and functional specification definition
   B. Terminal-System-Service information interface API for AI reliability authentication
4. Activities to establish standards for AI reliability vertical convergence technology
   A. AI convergence technology such as intelligent terminal authentication and smart monitoring through AI reliability certification and intelligent IoT platform convergence
   B. AI convergence technology such as intelligent content conversion, creation, and recommendation through intelligent content convergence
   C. AI convergence technology such as intelligent network allocation, quality adjustment, and monitoring through intelligent network convergence
5. Activities to establish standards for AI reliability certification technology
   A. AI reliability certification framework, standard conformance, interoperability test verification technology
6. Establishment of new committee in ISO/IEC JTC1, oneM2M and advance to the chairmanship

Hence, the outcome of the stage 1 will be as follows:

1. AI reliability research technical report: AI reliability research work item/technology report for their proposal, establishment, and adoption
2. AI Reliability Technology Standards: Establishment of AI reliability common standards/vertical convergence standards/certification standards
3. Establishment of committee and advance to chairpersonship through AI reliability technology standard enactment activity

4.2 Stage 2: Establishment of AI reliability certification global program

1. Designing and building AI reliability certification program
   A. Designing and building tools for AI reliability certification exam
   B. Development of test cases for AI reliability certification test

C. Seminars and training related to AI reliability certification test
D. Reliability certification program for Intelligent IoT platform convergence AI, launched through cooperation with IEC/ISO JTC1, oneM2M, and GCF (Note that GCF is Global Certification Forum.)
E. Establishment of related committee and appointment of chairmanship

2. Dissemination of AI reliability certification program
   A. Hosting a global InterOp event through the participation of AI-based product and service companies
   B. Introduction of open-source software to AI reliability certification test: open source test case and training of open source experts
3. AI reliability certification international standard seminar held
4. Establishment of TAIPP global AI reliability certification program
   A. Consensus derivation activities of related existing international standards organizations (JTC1/oneM2M)
   B. Governance maintenance activities within the organization
   C. Integrated and fast propagation activity in regional standard organizations (ETSI, TTA, etc.), which is a characteristic of Partnership Project (example: 3GPP, oneM2M)

5. Conclusion

In this paper, we propose the AI certification standardization activities in two stages. In the stage 1, as in the general standard activity procedure, we review the scope and possibility of standardization through AI reliability technology research targeting international standards organizations, we establish the AI reliability technology standard, and then we establish AI reliability verification for the feasibility of the AI reliability technology and certification standards. In the stage 2, based on the standard technical specifications established in the stage 1, we establish AI reliability certification program for verification of products, systems and services to which standard technologies are applied. Along with the establishment of the AI reliability certification system, a global InterOp (Interoperability test) event, an AI reliability certification international standard meetings and seminars are to be held for the spread of AI reliability certification. Finally, TAIPP (Trustworthy AI Partnership Project) will be established through the participation of relevant standards organizations and industries to overall maintain and develop standards and standards certification programs to ensure the governance of AI reliability certification standards.

Acknowledgement

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by ICT R&D program of MSIT/IITP [2021-0-00193, Development of photorealistic digital human creation and 30fps realistic rendering technology].

References

[1] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple Black-box Adversarial Attacks," in *Proc. ICML 2019,* Jun. 9-15, 2019.

[2] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," in *Proc. NIPS 2018,* Dec. 3-8, 2018.

[3] Z. Huang, and T. Zhang, "Black-Box Adversarial Attack with Transferable Model-based Embedding," in *Proc. ICLR 2020,* Apr. 26-May 1, 2020.

[4] H. Kwon, H. Yoon, and K.-W. Park, "Selective Poisoning Attack on Deep Neural Network to Induce Fine-Grained Recognition Error," in *Proc. IEEE 2nd International Conference on Artificial Intelligence and Knowledge Engineering (AIKE),* pp. 136-139, Jun. 3-5, 2019.

[5] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified Defenses against Adversarial Examples," in *Proc. ICLR 2018,* Apr. 30-May 3, 2018.

[6] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," in *Proc. NIPS 2018,* Dec. 3-8, 2018.

[7] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," in *Proc. 23rd International Conference on Artificial Intelligence and Statistics (AISTATS),* Apr. 13- 15, 2021.

[8] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in *Proc. ICML 2018,* Jul 10-15, 2018.

[9] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in *Proc. NeurIPS 2019,* Dec. 8-14, 2019.