SafeAI-Kit: A Software Toolbox to Evaluate AI Systems with a Focus on Uncertainty Quantification

Dominik Eisl
Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Bastian Bernhardt
Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Lukas Höhndorf
Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Rafal Kulaga
Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Abstract

In the course of the practitioner track, the IABG toolbox safeAI-kit is presented with a focus on uncertainty quantification in machine learning. The safeAI-kit consists of five sub-modules that provide analyses for performance, robustness, dataset, explainability, and uncertainty. The development of these sub-modules take ongoing standardization activities into account.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases safeAI-kit, Evaluation of AI Systems, Uncertainty Quantification

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.6

Category Practitioner Track

1 Introduction

In recent years, Machine Learning (ML) models have become immensely powerful and are used in a wide range of applications and domains. The rise of large language models (LLMs) has led to the popularization of Artificial Intelligence (AI) in society and boosted the already high interest of the industry. Deep learning models are deployed to estimate depth in 2D images [4], to detect abnormalities in medical images [2], or to convert speech to text, often with impressive results. But regardless of the AI task at hand, the number of parameters, or the used architecture, none of the AI models are perfect. Incorrect predictions and mistakes in outputs generated by the AI are inevitable. Given their imperfections, thorough testing and validation of AI models are crucial steps towards deploying them reliably and safely. In the following, we provide insights into the safeAI-kit, which is a toolbox for AI evaluation that offers methods for dataset analysis, performance and robustness evaluation, uncertainty quantification, and explainability examination. In addition, we focus on uncertainty quantification in machine learning.

Over the past years, we have dedicated ourselves to the development of solutions for evaluating and safeguarding AI systems. We combine state-of-the-art AI research, standardization, and regulation with the vast experience of IABG in testing, analyses, and certification processes. By actively participating in AI standardization committees on national, European, and international levels, we support the development of new standards and guidelines to increase the benefits of AI systems.

2 IABG's safeAI-kit

To conduct practical AI assessments, we identified the need of a testing tool to support future conformity assessments. The IABG's safeAI-kit is a software toolbox developed to support the evaluation of AI models and datasets and provides a comprehensive analysis with respect to five dimensions. These encompass dataset analysis, performance and robustness evaluation, uncertainty quantification, and explainability examination, all aiming at providing a thorough understanding of an AI model's behavior and capabilities. The resulting evaluation report presents detailed insights into the model's strengths, weaknesses, and overall capabilities, empowering the implementation of safety measures and streamline future audits.

In addition, IABG contributes to the current development of DIN SPEC 92006 "Artificial Intelligence - Requirements for AI testing tools" and aligns the safeAI-kit development with all the concepts and methodology described therein.

3 Uncertainty Quantification in Machine Learning

The real world is complex, chaotic, dynamically changing, and thus difficult to represent in a training set, from which models gain knowledge. Uncertainty is, therefore, inherent in the model's operation. For humans it is very natural to express uncertainty when faced with a new situation or a difficult question. We use phrases like "maybe", "probably" or "I don't know". Analogously, the goal of Uncertainty Quantification (UQ) in ML is to enable the models to signal whether they are confident about the provided output or, on the contrary, that they "don't know" and are in fact guessing.

Hence, uncertainty quantification is an important building block of AI safety which is also vital for various other ML techniques such as active learning. Calibrated uncertainty measures can improve decision making and trustworthiness during operation and therefore offer a step forward beyond offline performance evaluation. Attention and awareness of UQ is growing in scientific and standardization communities as well as industry. As the field of UQ in ML becomes more technically mature and witnesses growing adoption in mission-critical ML tasks, the importance of UQ for safety and certification of AI systems will rise which will help to establish a strong base for the trustworthiness of AI in our society.

Uncertainty Quantification has also already been investigated in standardization. Initiated by IABG and Fraunhofer IAIS and developed by a consortium of experts, DIN SPEC 92005 "Uncertainty quantification in machine learning" [1] is a standardization document which aims to support stakeholders in adopting UQ in ML. The document defines important terms related to uncertainty and provides an overview of UQ applications, approaches, and properties. The core part of [1] is a set of recommendations and requirements for incorporation of UQ in ML. These guidelines aim to help developers navigate through the field of UQ in ML and support them in ensuring that UQ is applied correctly.

4 Conclusion

As the use of AI in safety-critical applications is increasing, independent assessments of such AI systems are essential, particularly considering strict requirements imposed on them by forthcoming legal frameworks, such as the European AI Act [3]. Our goal with the safeAI-kit is to not only contribute to the testing of AI systems but also to support future AI conformity assessment procedures, in alignment with both, legal frameworks and technical standards. The safeAI-kit is continuously evolving to address the challenges arising from limited practical experience in evaluating AI systems for various uses cases and applications. The adoption of

UQ in ML will play a central role for ensuring that ML models are solving complex tasks in a safe and trustworthy manner. Moreover, other dimensions we address with the safeAI-kit – such as performance, robustness, explainability, and dataset analysis – are equally important and must be considered when assessing AI systems. To conclude, we consider AI assessments and audits as a vital step towards enabling robust, reliable, and trustworthy AI systems.

References

1 "DIN SPEC 92005:2024-03, Künstliche Intelligenz - Quantifizierung von Unsicherheiten im Maschinellen Lernen; Text Englisch". Technical report, DIN Media GmbH, Berlin, 2024. URL: https://www.dinmedia.de/en/technical-rule/din-spec-92005/376619718.

2 Minliang He, Xuming Wang, and Yijun Zhao. A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs. *Scientific Reports*, 11, 2021. URL: https://api.semanticscholar.org/CorpusID:233427277.

3 The European Parliament and the Council of the European Union. "REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (Artificial Intelligence Act)", 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

4 Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. arXiv:2401.10891, doi:10.48550/arXiv.2401.10891.