Reliable and Explainable AI in Trieste

Emanuele Ballarin[1], Luca Bortolussi[1], Francesca Cairoli[1], Chiara Gallese[1], Laura Nenzi[1] and Gaia Saveri[1,2]

[1]Department of Mathematics and Geoscience, University of Trieste, Italy
[2]Department of Computer Science, University of Pisa, Italy

Abstract

This paper summarizes the activity in the area of Reliable and Explainable AI carried out at the University of Trieste. The main topics are: monitoring and verification of stochastic systems, embeddings of logical formulae using Graph Neural Networks, formal methods for explainable AI, adversarial robustness, right to interpretability, ethical assessment of data sets and AI auditing.

1. Introduction

The research activity in the area of Reliable and Explainable Artificial Intelligence (AI) at the University of Trieste is mainly concerned with the following topics:

- Monitoring and Verification of Stochastic Systems;
- Embeddings of Logical Formulae; using Graph Neural Networks;
- Formal Methods for Explainable AI;
- Adversarial Robustness;
- Right to interpretability;
- Ethical assessment of data sets;
- AI auditing.

In the following, a brief description of each of the aforementioned problems is provided.

2. Research Topics

2.1. Monitoring and Verification of Stochastic Systems

Monitoring and parametric verification of stochastic systems are extremely challenging and computationally prohibitive. We introduce efficient and reliable approaches to approximate the two aforementioned problems.

Predictive Monitoring. Predictive monitoring (PM) deals with predicting at runtime the satisfaction of a desired property from the current system's state. PM methods need to be efficient to enable timely interventions against predicted violations, while providing correctness guarantees. We introduce *quantitative predictive monitoring (QPM)* [1], a PM method to support stochastic processes and rich specifications given in Signal Temporal Logic (STL). QPM derives prediction intervals that are highly efficient to compute and with probabilistic guarantees, in that the intervals cover with arbitrary probability the STL robustness values relative to the stochastic evolution of the system. To do so, we take a machine-learning approach and leverage recent advances in conformal inference for quantile regression, thereby avoiding expensive Monte Carlo simulations at runtime to estimate the intervals. We also show how our monitors can be combined in a compositional manner to handle composite formulas, without retraining the predictors or sacrificing the guarantees.

Parametric Verification. Parametric verification of linear temporal properties for stochastic models requires to compute the satisfaction probability of a certain property as a function of the parameters of the model. Smoothed model checking (smMC) [2] infers the satisfaction function over the entire parameter space from a limited set of observations obtained via simulation. As observations are costly and noisy, smMC leverages the power of Bayesian learning providing accurate reconstructions with statistically sound quantification of the uncertainty. We introduce Stochastic Variational Smoothed Model Checking (SV-smMC) [3], which exploits stochastic variational inference (SVI) to approximate the posterior distribution of the smMC problem. The strength and flexibility of SVI, a stochastic gradient-based optimization making inference easily parallelizable and enabling GPU acceleration, make SV-smMC applicable both to Gaussian Processes (GP) and Bayesian Neural Networks (BNN). SV-smMC extends the smMC framework by greatly improving scalability to higher dimensionality of parameter spaces and larger training datasets, thus overcoming the well-known limits of GP. Additionally, we combine the Bayesian quantification of uncertainty of SV-smMC with the Inductive Conformal Predictions framework to provide probabilistically approximately correct point-specific error estimates, with

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3486/26.pdf

statistical guarantees over the coverage of the predictive error.

2.2. Formal Methods for Explainable AI

Over the past few years, part of our research has focused on exploring the application of formal methods to enhance the explainability of machine learning components. Cyber-Physical Systems (CPS) are at high risk of cyber-attacks. A possibility to detect anomalies is monitoring CPS sensor and actuator data using Signal Temporal Logic (STL) formulas. In [4], we propose a one-shot algorithm to learn a set of STL formulas from a data set of only regular behaviors. The algorithm learns an ensemble of STL formulas using a Grammar-Guided Genetic Programming (G3P) algorithm. The ensemble can then be used to detect anomalous behaviors. Testing on real-world datasets shows that the proposed one-shot algorithm provides effective detection performance.

In [5], we exploit ML for predicting the monitoring result of STL formulas. We introduce a similarity function (as a kernel function) on the semantics space of STL formulae. The approach avoids the need to convert formulae into vectors of numbers explicitly. The proposed method is demonstrated to effectively predict the satisfaction of STL formulae on stochastic processes, with high precision and computational efficiency. Furthermore, the effectiveness of the method is supported by a theoretically sound PAC guarantee.

2.3. Embeddings of Logical Formulae using Graph Neural Networks

Logic is the main formal language to perform automated reasoning, and it is further a human-interpretable language, at least for small formulae. Learning and optimising logic requirements and rules has always been an important problem in Artificial Intelligence. State of the art Machine Learning (ML) approaches are mostly based on gradient descent optimisation in continuous spaces, while learning logic is framed in the discrete syntactic space of formulae. Using continuous optimisation to learn logic properties is a challenging problem, requiring to embed formulae in a continuous space in a meaningful way, i.e. preserving the semantics. Approaches like [5] are able to construct effective semantic-preserving embeddings via kernel methods (for linear temporal logic), but the map they define is not invertible. We address this problem, learning how to invert such an embedding leveraging deep architectures based on the Graph Variational Autoencoder framework, proposing a novel model specifically designed for this setting.

2.4. Adversarial Robustness

In an ongoing research endeavour, we are developing a novel technique to improve adversarial robustness of neural image classifiers, broadly inspired by the process of memory recollection in humans and synergistically complementing the first-line approach of adversarial training. Specifically, we use the representation produced in selected layers of an adversarially-pretrained deep neural network in response to clean and perturbed inputs to condition an adversarial purification generative model, whose output is finally classified conventionally. Preliminary results – to be soon submitted to a major conference in the field – show significant improvement in robustness, as evaluated by [6] on different classifier architectures and standardised image classification datasets, with acceptable clean accuracy degradation.

In a further development, such approach is employed to identify the layers of a deep learning model mostly contributing to its adversarial vulnerability, with the goal of producing partially-Bayesian neural networks whose robustness-enhancing effects (as analysed in previous works, *e.g.* [7]) are focused where they are needed most, bounding the computational burden that follows.

Finally, a flipped perspective will guide the development of a new assessment framework for surrogate deep Bayesian models, exploiting the striking characteristic robustness properties of proper Bayesian neural networks, and how they are mirrored in surrogates.

2.5. Right to interpretability, ethical assessment of data sets, and AI auditing

We study the legal, ethical and societal issues related to the use of AI system. In particular, taking into account the systematic interpretation of the EU and international legal framework surrounding high-risk systems, we research about the possibility that interpretability, as defined by Rudin, may be used as a standard in in applications that may have a significant impact on citizens' life (and even in other sensitive fields) and that black boxes should only be used in situations where it is possible to make a decision by evaluating factors other than the AI output. The very possibility of expressing informed consent and challenging the decision made on the basis of an automated decision-making system might be excluded by the opacity and complexity of black boxes. The lack of technical interpretability may prevent the exercise of many fundamental rights, such as the right to a fair trial, to self-determination, to non-discrimination, and more. We argue that a "right to technical interpretability" should, therefore, be theorized at the European level, being considered a fundamental right, and embedded in the AI Act proposal [8].

Within the EU AI Act proposal, major attention is given to the assessment of data sets, that are required to be unbiased. In our research, we are piloting new techniques to assess data sets' fairness and we aim at providing guidelines to comply with the new AI Act data set compliance framework.

We are also exploring the field of AI Auditing in light of the applicable EU and international laws and regulations, such as Convention 108 plus, GDPR, the Data Act, the AI Act, the AI liability Directive, the Cyber-resilience Act, and such. Our aim is to contribute to research on Trustworthy AI by exploring how to apply abstract principles in practice, embedding them in the whole AI developing lifecycle.

Finally, we research about the social issues related to AI systems. Generative models such as Midjourney, Dall-e, Stable Diffusion, and Chat GPT have been increasingly used to create new content on the basis of a user's prompt, producing impressive images and texts; however, when copyrighted materials and personal data are used to train the model, many ethical and legal issues arise. In fact, such large data set employed for the AI training are collected by infringing a number of different laws and contracts, and, most importantly, without the consent of the interested parties, leading to detrimental effects towards citizens and the society as a whole.

References

[1] L. Bortolussi, F. Cairoli, N. Paoletti, Conformal quantitative predictive monitoring of stl requirements for stochastic processes, in: 26th ACM International Conference on Hybrid Systems: Computation and Control, 2023.

[2] L. Bortolussi, D. Milios, G. Sanguinetti, Smoothed model checking for uncertain continuous-time markov chains, Information and Computation 247 (2016) 235–253.

[3] L. Bortolussi, F. Cairoli, G. Carbone, P. Pulcini, Stochastic variational smoothed model checking, arXiv preprint arXiv:2205.05398 (2022).

[4] P. Indri, A. Bartoli, E. Medvet, L. Nenzi, One-shot learning of ensembles of temporal logic formulas for anomaly detection in cyber-physical systems, in: E. Medvet, G. L. Pappa, B. Xue (Eds.), Genetic Programming - 25th European Conference, EuroGP 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings, volume 13223 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 34–50. URL: https://doi.org/10.1007/978-3-031-02056-8_3. doi:10.1007/978-3-031-02056-8_3.

[5] L. Bortolussi, G. M. Gallo, J. Kretínský, L. Nenzi, Learning model checking and the kernel trick for signal temporal logic on stochastic processes, in: D. Fisman, G. Rosu (Eds.), Tools and Algorithms for the Construction and Analysis of Systems - 28th International Conference, TACAS 2022, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2022, Munich, Germany, April 2-7, 2022, Proceedings, Part I, volume 13243 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 281–300. URL: https://doi.org/10.1007/978-3-030-99524-9_15. doi:10.1007/978-3-030-99524-9_15.

[6] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, RobustBench: a standardized adversarial robustness benchmark, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021. URL: https://openreview.net/forum?id=SSKZPJCt7B.

[7] G. Carbone, M. Wicker, L. Laurenti, A. Patane', L. Bortolussi, G. Sanguinetti, Robustness of bayesian neural networks to gradient-based attacks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 15602–15613. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/b3f61131b6eceeb2b14835fa648a48ff-Paper.pdf.

[8] C. Gallese, The AI act proposal: a new right to technical interpretability?, arXiv preprint (2023).