On Audit and Certification of Machine Learning Systems

Dmitry Namiot

Lomonosov Moscow State University
dnamiot@gmail.com

IT Congress 2023

Outline

- Software for critical applications (avionics, etc.) - is certified to prove functionality.
- Inference-stage machine learning systems are also software.
- Accordingly, they must also be certified.
- Is it possible?
- What practically can be done in this direction?

On certification

- For software in critical areas, the developer must ensure its safety and demonstrate its functionality.
- The term confirmation of conformity is also used - that is, a demonstration that the product has all the necessary properties (the software implements all the necessary functions).
- Software certification confirms, first of all, its functional safety (correct and complete implementation of the declared functions).
- The documents that describe such certification may vary from area to area. In aviation this is the DO-178C standard, in nuclear energy - IEC 60880, etc.

On certification

- The documents are different, but the software quality control methods proposed in them are, in fact, almost the same.
- Software quality control always includes functional testing, which consists of checking the compliance of real and declared functionality, structural testing (coverage of execution routes by tests - full coverage is usually required), unit testing, analysis of code metrics, control of data flow, and data connectivity.
- Certification can also affect the development process (working with requirements, configuration management, etc.)

On certification

- Machine learning systems at the inference (operation) stage are, generally speaking, “regular” software that should not be considered as something separate from the overall process.
- The actual inference module can be portable (standard), the trained model is represented by some file (a parameter for the inference module).
- So, machine learning systems must be certified in the same way as all other software. And this is where the problem arises.
- The first is the non-deterministic nature of the responses and problems with reproducibility.
- Secondly, there is the presence of adversarial attacks

Secure/Safe ML (MIT)

- I. Thou shall not train on data you don’t trust (because of data poisoning)
- II. Thou shall not let anyone use your model (or observe its outputs) unless you completely trust them (because of model stealing and black box attacks)
- III. Thou shall not fully trust the predictions of your model (because of adversarial examples)

On adversarial attacks

- On adversarial attacks as per NIST:

- Poisoning
- Evasion
- Attacks on intellectual property
- Evasion attacks are inevitable !
- Classically: small changes to correct input

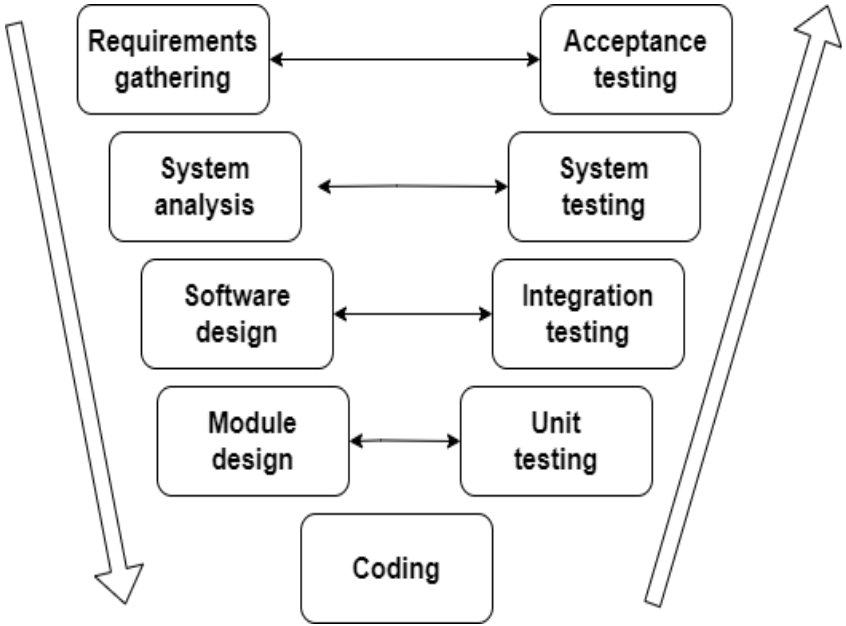
On ML robustness

- Classically, for machine learning systems, robustness is defined as the independence of the output of the system from small changes in the input data.
- Given an input *x* and a model of interest *f*, we want the model prediction to remain the same for all inputs *x′* in a neighborhood of *x*, where the neighborhood is defined by some distance function *δ* and some maximum distance *Δ*: $\delta(x, x') \leq \Delta \; f(x)=f(x')$
- In practice for critical systems: the model maintains the performance shown in the training phase for all inputs

ML robustness and security

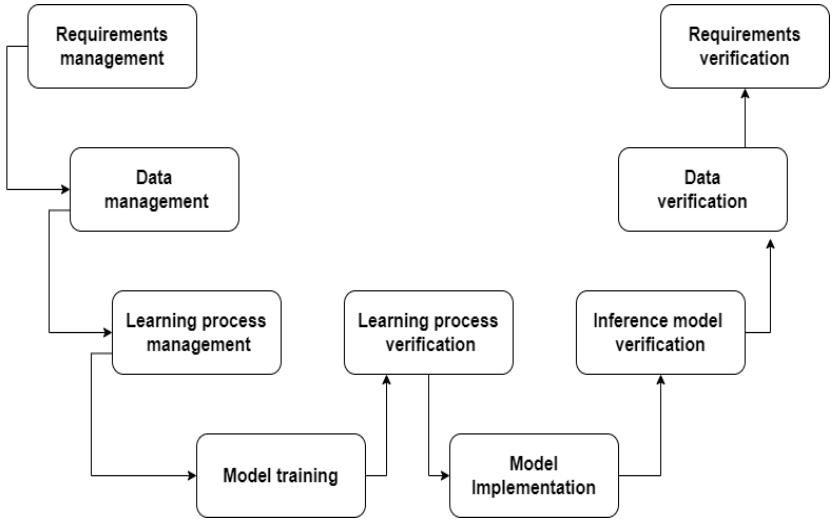
- robustness != security
- The definition for robustness does not say anything about the correct operation of the system
- That is, there may well be a robust system that produces incorrect results. And these incorrect results remain so for small perturbations of the initial data.
- Hence, robustness by itself cannot be indicative of software safety. Safety (security) is a property of a system that includes a machine learning model. With regard to machine learning systems, security is used as a synonym for trust in the results of work

Software assurance



- There is a classical V-model of software development. Two test directions:
- Verification – are we building the product correctly?
- Validation - is the correct product built?
- Each level has a corresponding set of tests.

ML model



- For machine learning systems, each step exists on its own.
- So, the problem is reduced to a sequence of deterministic steps (at each step, some deterministic result is obtained).
- It turns out that the question of a possible shift of data generally falls out of consideration.

The main inconsistencie

- deterministic approach in certification of software systems against non-deterministic ML models
- code coverage (it is based on V-model - why is this line in the code?)
- data coverage. A standard approach in ML is point-wise robustness. Certification for ML models is a study of robustness in a limited range of modifications of correct data. E.g. 35.42% certified accuracy on MNIST under perturbation ϵ= 8/255. Here is the data set (correct images) and the limits of change for the pixel (8)

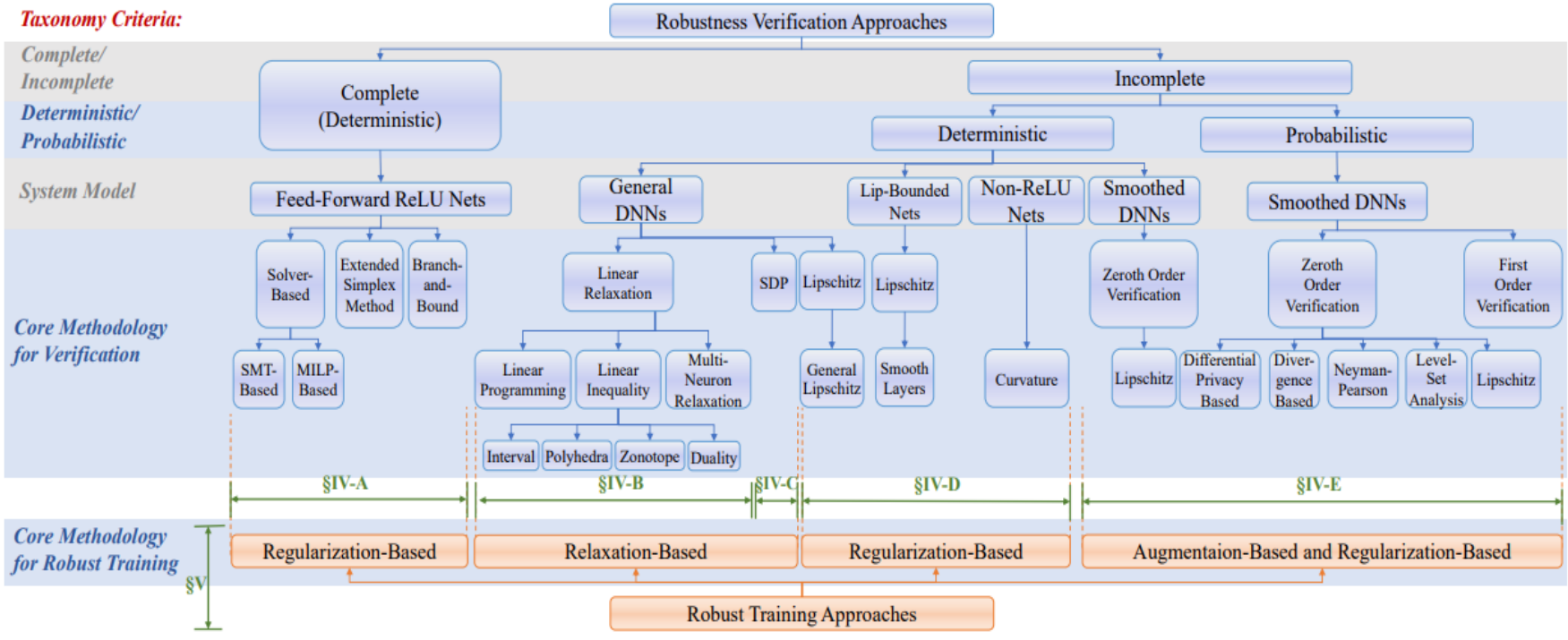
Certified robustness

- We can distinguish two approaches. Firstly, it is a formal verification of machine learning models.
- This is a working approach with one big problem - scalability. Verification of models is reduced to logical formulas, or to systems of linear equations, and with an increase in the number of parameters (and now it is no longer millions of parameters), the task becomes unsolvable.
- Robustness testing approaches aim to assess the robustness of neural networks by providing a theoretically validated lower bound on robustness under certain perturbation constraints.

Certified robustness

- Complete validation and incomplete validation: when the check method outputs "not checked" for a given x_0, if it is guaranteed that there is an adversarial instance of x around x_0, we call this a complete check, otherwise, it is an incomplete check.
- Deterministic verification and probabilistic verification: when given inputs are not resistant to attack, deterministic testing guarantees the output "not tested", and probabilistic testing guarantees the output "not tested" with a certain probability (for example, 99.9%), where the randomness does not depend on the input data.

Certified robustness



Certified robustness

Dataset	Method	FLOPs	Test	Robust	Certified
MNIST ($\epsilon = 0.3$)	Group Sort (Anil et al., 2019)	2.9M	97.0	34.0	2.0
	COLT (Balunovic & Vechev, 2020)	4.9M	97.3	-	85.7
	IBP (Gowal et al., 2018)	114M	97.88	93.22	91.79
	CROWN-IBP (Zhang et al., 2020b)	114M	98.18	93.95	92.98
	ℓ_∞-dist Net	82.7M	98.54	94.71	92.64
	ℓ_∞-dist Net+MLP	85.3M	**98.56**	**95.28**	**93.09**
Fashion MNIST ($\epsilon = 0.1$)	CAP (Wong & Kolter, 2018)	0.41M	78.27	68.37	65.47
	IBP (Gowal et al., 2018)	114M	84.12	80.58	77.67
	CROWN-IBP (Zhang et al., 2020b)	114M	84.31	80.22	78.01
	ℓ_∞-dist Net	82.7M	**87.91**	79.64	77.48
	ℓ_∞-dist Net+MLP	85.3M	**87.91**	**80.89**	**79.23**
CIFAR-10 ($\epsilon = 8/255$)	PVT (Dvijotham et al., 2018a)	2.4M	48.64	32.72	26.67
	DiffAI (Mirman et al., 2019)	96.3M	40.2	-	23.2
	COLT (Balunovic & Vechev, 2020)	6.9M	51.7	-	27.5
	IBP (Gowal et al., 2018)	151M	50.99	31.27	29.19
	CROWN-IBP (Zhang et al., 2020b)	151M	45.98	34.58	33.06
	CROWN-IBP (loss fusion) (Xu et al., 2020a)	151M	46.29	35.69	33.38
	ℓ_∞-dist Net	121M	**56.80**	**37.46**	33.30
	ℓ_∞-dist Net+MLP	123M	50.80	37.06	**35.42**

EASA roadmap

- Level 1 – assistance to human:
  - Level 1A: Human augmentation
  - Level 1B: Human cognitive assistance in decision-making and action selection
- Level 2 – human-AI teaming
  - Level 2A: Human and AI-based system cooperation
  - Level 2B: Human and AI-based system colloboration

EASA roadmap

- Level 3 – advanced automation

  Level 3A: The AI-based system performs decisions and actions that are overridable by the human

  Level 3B: The AI-based system performs non-overridable decisions and actions (e.g. to support safety upon loss of human oversight)
- Certification of applications of the first level (human assistants) refers to 2025, and the last third level (non-cancellable actions) - to 2035-2050.

AI audit

- ML modeling: alternate ML approaches; reasons to justify chosen ML strategy; refining ML algorithms.
- AI project scope definition: constraints and other implementation approaches.
- Deployment and testing: methods used to deploy ML models; post deployment review; metric used to ensure accuracy of ML models.
- Data management: data sources and data consistency; data imputation and data standardization.
- Data monitoring: monitor model performance, drift, activities, and anomalies; compliance with law and regulatory standards; ethical and social responsibility.
- AI-audit is exactly the checklist for checking the availability of the necessary activities.

AI audit base

- EASA released a technical report, “EASA Concept Paper: First Useful Guidance for Level 1 Machine Learning Applications”, which describes best practices in the design and development of machine learning systems.
- The questionnaire is based on the Assessment List for Trustworthy AI (ALTAI) project , prepared by a group of AI HLEG (High-level expert group on artificial intelligence) experts working with the European Commission and the European AI Alliance.
- The questionnaire includes 7 main top-level sections:

AI audit base

- 1) Human Agency and Oversight
- 2) Technical Robustness and Safety
- 3) Privacy and Data Governance
- 4) Transparency
- 5) Diversity, Non-discrimination and Fairness
- 6) Societal and Environmental Well-being
- 7) Accountability

AI audit example

- Have you put measures in place to ensure the traceability of the AI system throughout its lifecycle?
- Have you taken steps to continually evaluate the quality of the input data to the AI system?
- Can you trace what data was used by the AI system to make specific decision(s), or recommendations?
- Can you trace which model, or which rules led to the decisions or recommendations of the AI system?
- Have you taken steps to continually evaluate the quality of the AI system's output?
- Have you implemented adequate logging techniques to record the decisions or recommendations of the AI system?

Q&A as conclusion

1) Do the adopted and planned regulations relate to the certification of machine learning systems? (the answer is no);

2) Are there enough robustness checks for secure systems? (no);

3) Is it currently possible to certify machine learning systems according to the same scheme as software in critical systems is certified? (no);

4) Certification of the robustness of machine learning models has very little to do with certification of software implementations of machine learning models (yes);

5) What is a necessary and feasible step towards certification of machine learning systems? (audit);

6) What could be the basis for an auditing standard? (EASA / ALTAI)