Reliable and Trustworthy Artificial Intelligence

Lecture 13: Summary and future directions

Martin Vechev

ETH Zurich

Fall 2022

RTAI: what was advertised

Robustness

attacks and defenses, certification (relaxations, branch and bound, certified training, smoothing), logic + deep learning

Privacy

attacks, differential privacy, secure synthetic data, data minimization, federated learning vulnerabilities

Fairness/Bias

individual fairness, group fairness, methods for building fair systems for tabular, NLP and visual data

Common theme: provable mathematical guarantees for all of the above

RTAI: Exam (Very) Rough Structure

1. Multiple choice (yes/no questions mostly from topics not covered below)
2. Attacks and defenses
3. Box/DeepPoly/BnB/Certified Training/MILP
4. DP & Randomized Smoothing
5. Federated / DP for ML
6. Individual/Group fairness, logic/loss

Open Research Problems

1. **Randomized smoothing:** how tight is it? Lower vs. upper bounds?
2. **Randomized smoothing:** how can we fine tune with it on custom data?
3. **Randomized smoothing:** can we define constraints over the input such that each partition is smoothed separately?
4. **Proof Transfer:** can we do proof transfer with essentially combination of multi-neuron constraints and KKT?
5. **Certification:** custom relaxations beyond Box that lift the SABR certified training method to richer relaxations?
6. **Certified training:** training with multi-neuron constraints
7. **Differential Privacy:** can we define a language to express privacy policies and synthesize custom noise?
8. **Differential Privacy:** non-membership inference (check ove that some data is not used for training)
9. **Federated learning:** extending Fed-Avg to larger datasets and systems, closer to practice.
10. **Federated learning:** attacks on graph neural networks and data
11. **Private and Reliable Inference:** Randomized smoothing with secure multi-party computation.
12. **Formalized regulations:** formalize latest regulations (e.g., linkability, etc.) and devise attacks for these
13. **Blind spots:** blind spots in NLP and vision models
14. **Group Fairness:** generalization of FARE + trade-offs
15. **Group Fairness:** FRL benchmark
16. **Group Fairness:** Learning the groups for debugging the models
17. **Large Language models [a range of topics]:** languages, filters, de-biasing, synthesis, data extraction, decomposition, etc.
18. **…**

Semester Thesis, Research, M.Sc. Thesis

Many students who took the course published results in top AI/ML conferences **as part of their M.Sc./semester thesis**.

If interested in doing research in this space in Spring 2023 or later, let me know and we can discuss.

Hope you had fun and happy holidays to all ☺