Rule-based Safety Evidence for Neural Networks

- Tewodros A. Beyene
- Amit Sahu

Outline

- Introduction
- Rule Extraction
- Rule-based Safety Evidence
- Illustration
- Results
- Conclusion and Future Work

Introduction

- Safety-critical systems are often subject to a rigorous safety certification process (DO-178C, ISO26262)
- In general, we need evidence for safety certification as information or artefacts that contribute to developing confidence in the safe operation of a system
- There have been efforts to extend safety assurance approaches for traditional software to AI-based systems, such as self-driving cars and other autonomous systems.
- These face two challenges:
  - Efficient analysis, testing, and verification methods
  - Finding artefacts that develop confidence in the safe operation of the AI system
- In this position paper, our focus was on second challenge by using rule extraction techniques

Rule Extraction

- Approaches
  - Pedagogical: treat network as black-box
  - Decompositional: split the network at neuron level
- Types of Extracted Rules: If-then, M-of-N, Decision Tree, fuzzy rules, first-order rules
- Rule Extraction for DNNs
  - Generated layer-by-layer
  - High memory usage and computation time
  - Rule pruning and network pruning

Rule-based Safety Evidence

Rules as Evidence Artefacts

- Diverse perception/view of rules: coverage over features, robustness to noise
- Applicability of extraction techniques in diverse architectures: locally in federated learning, networks that facilitate rule extraction
- Amenability of rules for quality assessment: criteria like accuracy, fidelity, and consistency
- Diverse expressive power of rules: for example, decision tree can express hierarchical properties of an NN as compared to if-then rules

Rule-based Safety Evidence

Algorithmic complexity of rule extraction

- Scaling of decompositional approaches
- Various heuristics to limit space exploration and achieve tractability for real world problems
- Complexity of extracted rules: number of extracted rules, number of antecedents per rule
- Solutions for reducing the size of generated rulesets: network pruning, rule pruning, rule abstraction, distillation, formal methods: partial order reductions, priority synthesis

Illustration

Methodology

- Select Dataset: MNIST
- Train NN models with the goal of achieving accuracy and specific properties: robust to adversarial noise
- Extract rules from the trained NN models
- Use these rules as safety evidence for the targeted properties

Illustration

NN Models

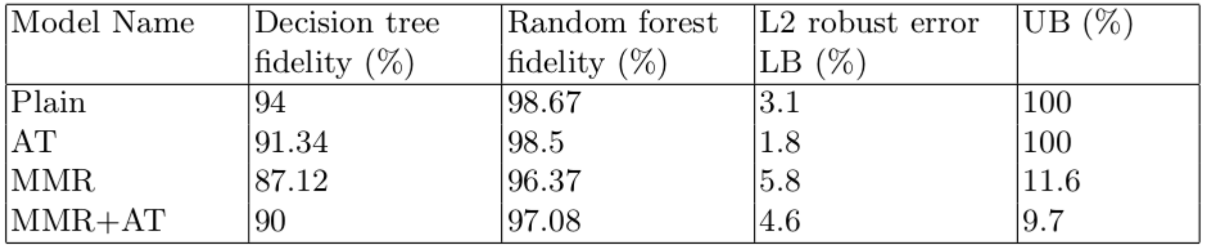
- Plain: trained with basic backpropagation algorithm
- Adversarial Training (AT): trained with adversarial images to make the model robust to adversarial perturbation
- Maximization of Linear Regions(MMR): trained with MMR regularizer to increase the provable bound on effective adversarial noise
- MMR+AT: a hybrid approach that uses both MMR and AT approaches

Illustration

Techniques

- TREPAN: The algorithm uses an Oracle that can query and sample examples from the dataset. Each node is split based on the best fit that generates the highest fidelity with the NN model
- Surrogate Random Forest Model: In this algorithm, the NN model was taken as a black-box model generating feature-output combination. This combination is used to train a separate Random Forest Model from scratch.

Results



Model Name	Decision tree fidelity (%)	Random forest fidelity (%)	L2 robust error LB (%)	UB (%)
Plain	94	98.67	3.1	100
AT	91.34	98.5	1.8	100
MMR	87.12	96.37	5.8	11.6
MMR+AT	90	97.08	4.6	9.7

Table 1. Fidelity of models with different surrogate models and adversarial robustness

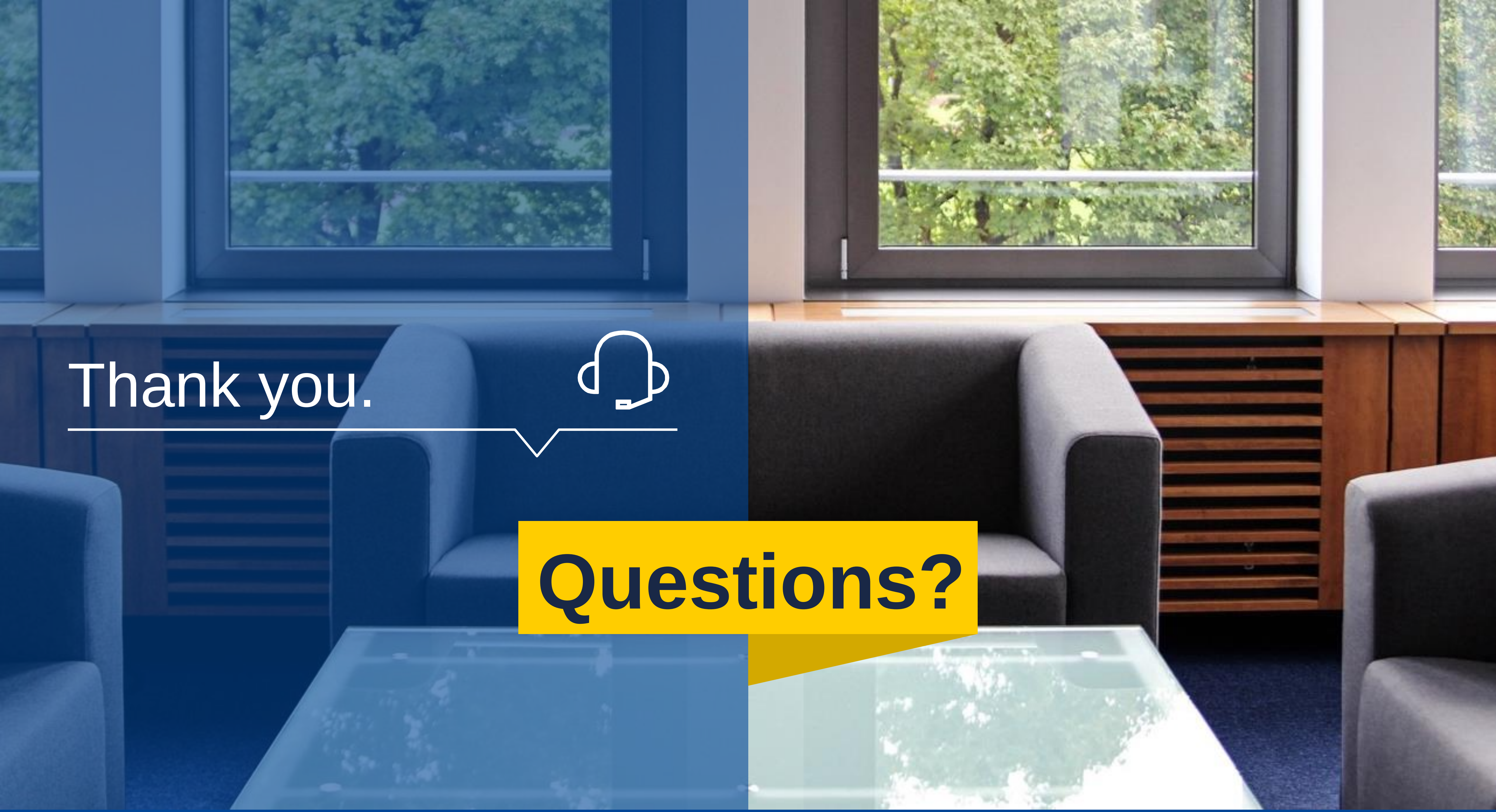
- Surrogate RF model had better fidelity scores than the surrogate Decision Tree attributed to higher complexity of RF model
- Models with robustness against adversarial property achieved lower fidelity than plain or simpler models

Conclusion and Future Work

- We have shown rules extracted from NN can be good evidence artefacts
- We observed that different safety properties can be defined, represented, and verified for critical systems using rules with high fidelity scores

Some properties that we will be working on for different datasets:

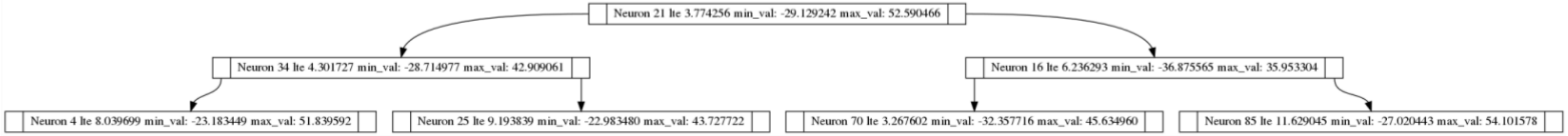
- Interpretation of the decision process
- Robustness to noise
- Tractable coverage of the decision tree as compared to coverage over neuron activations



Thank you.

Questions?

Surrogate Subtrees of the Models



Neuron 21 lte 3.774256 min_val: -29.129242 max_val: 52.590466

Neuron 34 lte 4.301727 min_val: -28.714977 max_val: 42.909061

Neuron 16 lte 6.236293 min_val: -36.875565 max_val: 35.953304

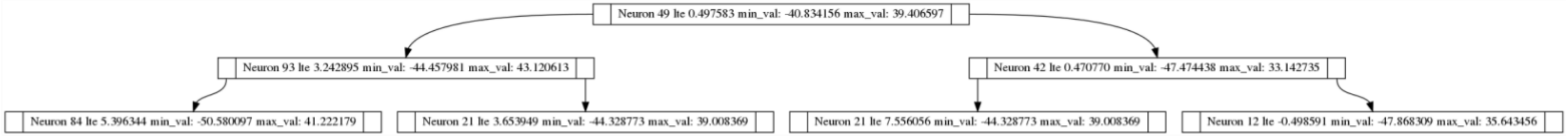
Neuron 4 lte 8.039699 min_val: -23.183449 max_val: 51.839592

Neuron 25 lte 9.193839 min_val: -22.983480 max_val: 43.727722

Neuron 70 lte 3.267602 min_val: -32.357716 max_val: 45.634960

Neuron 85 lte 11.629045 min_val: -27.020443 max_val: 54.101578

Surrogate Tree for MNIST plain model



Neuron 49 lte 0.497583 min_val: -40.834156 max_val: 39.406597

Neuron 93 lte 3.242895 min_val: -44.457981 max_val: 43.120613

Neuron 42 lte 0.470770 min_val: -47.474438 max_val: 33.142735

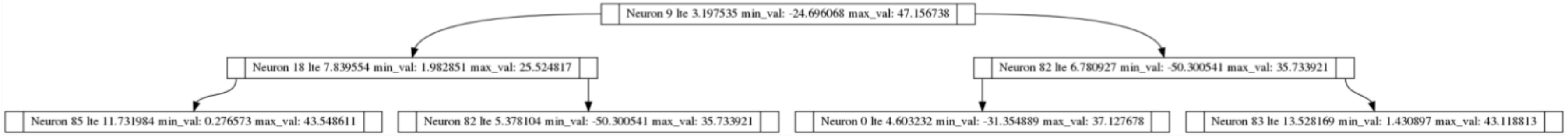
Neuron 84 lte 5.396344 min_val: -50.580097 max_val: 41.222179

Neuron 21 lte 3.653949 min_val: -44.328773 max_val: 39.008369

Neuron 21 lte 7.556056 min_val: -44.328773 max_val: 39.008369

Neuron 12 lte -0.498591 min_val: -47.868309 max_val: 35.643456

Surrogate Tree for MNIST AT model



Neuron 9 lte 3.197535 min_val: -24.696068 max_val: 47.156738

Neuron 18 lte 7.839554 min_val: 1.982851 max_val: 25.524817

Neuron 82 lte 6.780927 min_val: -50.300541 max_val: 35.733921

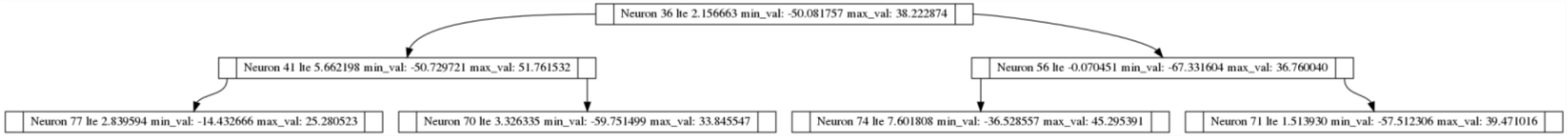
Neuron 85 lte 11.731984 min_val: 0.276573 max_val: 43.548611

Neuron 82 lte 5.378104 min_val: -50.300541 max_val: 35.733921

Neuron 0 lte 4.603232 min_val: -31.354889 max_val: 37.127678

Neuron 83 lte 13.528169 min_val: 1.430897 max_val: 43.118813

Surrogate Tree for MNIST MMR model



Neuron 36 lte 2.156663 min_val: -50.081757 max_val: 38.222874

Neuron 41 lte 5.662198 min_val: -50.729721 max_val: 51.761532

Neuron 56 lte -0.070451 min_val: -67.331604 max_val: 36.760040

Neuron 77 lte 2.839594 min_val: -14.432666 max_val: 25.280523

Neuron 70 lte 3.326335 min_val: -59.751499 max_val: 33.845547

Neuron 74 lte 7.601808 min_val: -36.528557 max_val: 45.295391

Neuron 71 lte 1.513930 min_val: -57.512306 max_val: 39.471016

Surrogate Tree for MNIST MMR+AT model