Formal Methods and Software Engineering in Certifying Deep Learning

AISafety2019@IJCAI2019
Xiaowei Huang, University of Liverpool
Outline

- Engineering AI Safety
- Safety Risk
- State-of-the-art
- Challenges and Gaps
Engineering AI Safety

- AI Safety Foundations
- Specification and Modelling
- Verification and Validation
- Runtime Monitoring and Enforcement
- Process Assurance and Certification
- Human-Machine Interaction
- Safety-related Ethics, Security and Privacy
Engineering AI Safety

Safety Risk

State-of-the-art

Challenges and Gaps
Safety Risk

- Robustness risk
- Generalisation risk
- Understanding risk
- Interaction risk
Robustness Risk

- Risk: a small perturbation on the input may lead to a significant difference in terms of the decision making.
Robustness Risk

...
Generalisation Risk

Risk: a change to scenario (environment, unimportant features, etc) leads to unexpected change in decision.
Understanding Risk

- e.g., why does this image represent a traffic sign of “go ahead or turn right”, instead of “go ahead”? 

- Risk: a decision is based on incorrect understanding about the input. This can easily lead to wrong decisions.
(Un-)reliability of a vehicle tracking system in wide area motion imagery (WAMI) where there are deep learning components.

- Risk: interaction with other components may introduce risks.
Engineering AI Safety

Safety Risk

State-of-the-art

Challenges and Gaps
State-of-the-art

- formal verification
  - constraint solving based methods
  - approximation methods
  - anytime methods
- engineering based methods
  - test coverage metric & test case generation

All for “Verification and Validation” and robustness risk, how about “specification and modelling”, “runtime monitoring and enforcement”, and ”process assurance and certification”?
General idea of constraint solving based methods

- Data collection & model construction & training
  - Safety requirement
  - Formula
  - Optimization or decision problem
  - Formula
  - Constraint solver
  - Satisfied
  - Unsatisfied
General idea of approximation methods

- Input region whose property is to be verified
- Safe output region
- Compute an over-approximation of the set
- Is this a subset relation?

Yes

No
General idea of anytime method
Coverage-based testing methods

- safety requirement
- test criteria
- test conditions
- coverage evaluation
  - above thresholds
  - lower than thresholds
- test case generation
- test cases
- oracle
- test statistics
Engineering AI Safety

Safety Risk

State-of-the-art

Challenges and Gaps
Challenges and Gaps

Specification and Modelling

▶ what is formal language (as Petri Nets, CASL, UML, etc in traditional software/hardware)?
▶ specification = training dataset?
Challenges and Gaps

- Verification and Validation
  - improved scalability for robustness verification
    - need a level of abstraction $\Rightarrow$ the use of domain-related information $\Rightarrow$ e.g., neuron $\rightarrow$ feature?
  - compositionality
  - verification for generalisation and understanding
Process Assurance and Certification

- not only for the final product but also for the development cycle: data collection, model construction, training, system integration, etc
- a successful assurance case: extension of safety argument method and goal structuring notation
Challenges and Gaps

► for both V&V and engineering based certification
  ► for real-time learning system, we need Runtime Monitoring and Enforcement for operational errors
  ► work with distributed learning system such as federated learning