Poster: Model checking RNNs with modal μ-calculus

Tatsuhiro Aoshima, Toshinori Usui
NTT Secure Platform Laboratories
{tatsuhiro.aoshima.md, toshinori.usui.rt}@hco.ntt.co.jp

***Abstract*—Machine learning models have been applied to many cyber-physical systems such as self-driving cars, robotics, and factory automation. However, it would be difficult to adapt them to more mission-critical systems, such as energy plants, because there is no safety guarantee. This poster presents the security of systems controlled using machine learning models, especially, Recurrent Neural Networks (RNNs). We propose a novel method for checking whether a given RNN satisfies a given specification, as abstractly interpreting the model with the constrained zonotopes. The specification is written in the modal μ-calculus containing many classical temporal logics such as Linear Temporal Logic and Computation Tree Logic.**

I. Introduction

Machine learning models have been applied to many cyber-physical systems such as self-driving cars, robotics, and factory automation. However, there is no guarantee for them to act safely, so it would be difficult to adapt them to more mission-critical systems such as energy plants. Attacks against these systems would seriously disrupt our society.

For example, if some control systems in a nuclear power plant are taken over by an attacker, the attacker could damage the plant and the neighbouring residents. Consider that a control rod in the plant keeps the temperature in the plant nearly constant and is controlled using an RNN. In this case, the developers would try to ensure that, for example, if the temperature x_T is greater than or equal to α (the threshold at which the control rod should start to work), then the position of the control rod o_P must be lower than or equal to γ (the threshold at which the control rod works) at some point in the future. This can be written formally in a modal μ-formula [2] as follows:

$$\nu x.(x_T \geq \alpha \rightarrow (\mu x.o_P \leq \gamma \vee \Box x)) \wedge \Box x$$

The pattern $\nu x.\psi \wedge \Box x$ means ψ is satisfied in any circumstance, and $\mu x.\psi' \vee \Box x$ means ψ' will be satisfied at some point in the future. This μ, ν operator directly corresponds to the corresponding model checking algorithm. Unlike DeepMind's method [3], it can specify over a period of time.

In this case, a model checking algorithm takes an RNN as a checked model and a modal μ-formula as a specification. It checks and outputs whether the given model satisfies the given specification. Then, if not so, it generates a counterexample that is an input to the RNN causing it to violate the specification.

We propose a novel method to make it possible to perform model checking of a given specification written in the modal μ-calculus on a given RNN. First, the algorithm (A) calculates the set of states of the RNN not satisfying the given specification (the semantic set) then (B) checks whether the given initial state is in it. If not so, (C) the RNN is concluded as satisfying the specification because, due to the construction of the semantic set, the initial state satisfies the specification. If so, the RNN is concluded as not satisfying the specification, and the algorithm calculates a counterexample with backpropagation using the calculation process of (A).

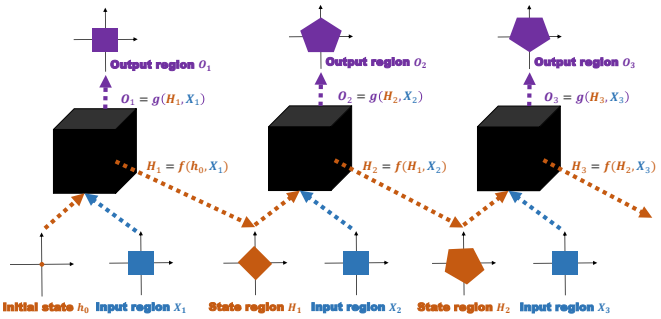


Fig. 1: Approximation of calculation of RNN with constrained zonotopes. Except for initial state $\boldsymbol{h}_0$, all input, state and output vectors are abstractly represented as set expressed as constrained zonotopes. Constrained zonotopes are closed under forward and backward computation of RNN, so set of states not satisfying given specification can be computed along with computation of RNN.

Technically, an RNN is expressed as a non-linear function composed of linear mapping layers and activation functions. Hence, to calculate the semantic set, our method interprets the model abstractly with the constrained zonotopes (see Fig. 1).

II. Technical Background

The safety of machine learning models is becoming an increasing concern. OpenAI released Safety Gym [4] to provide a framework for ensuring that machine learning models respect safety constraints. It can be used only in training new models and cannot be applied to trained models. Furthermore, it is not guaranteed that a model can satisfy the given safety constraints mathematically.

To mathematically guarantee safety, many model checking algorithms have been proposed. These only supports Finite State Machines (FSMs) [2] or PieceWise Affine (PWA) Continuous State Machines (CSMs) [1]. An RNN is a CSM

Model checking RNNs with modal μ-calculus

Tatsuhiro **Aoshima**, **T**oshinori **Usui** (NTT Secure Platform Laboratories)

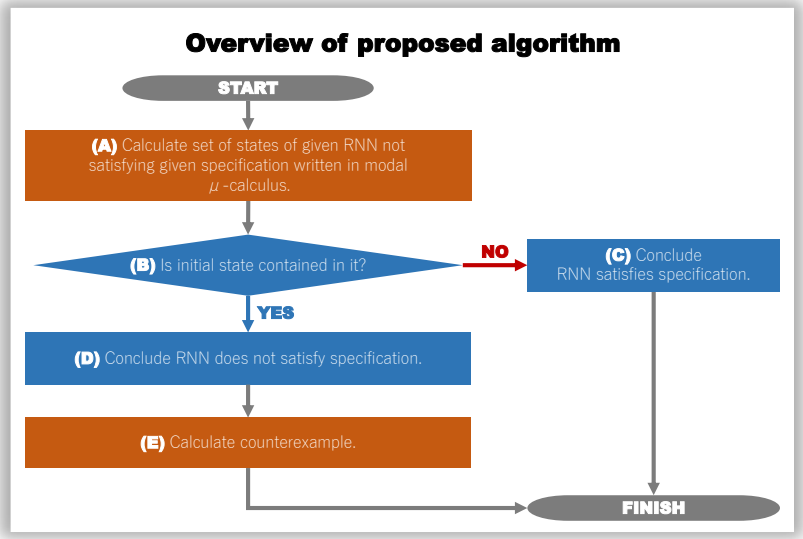
Machine Learning models have been **applied to cyber-physical systems** such as *self-driving cars*, *robotics*, and *factory automation*.

However, there is **no safety guarantee**, so, *attacks would seriously disrupt our society*.

We consider the **security of RNNs** as

(A) *abstractly interpreting* an RNN,
(B),(C),(D) checking whether an RNN satisfies a specification written in *modal μ-calculus*,
(E) generating *a possible attack pattern* if not so,

to check the safety *mathematically*.



Analogy of specification written in modal μ-calculus

Each *formula corresponds to a set of states*:

- $\Box p$ is a set of states satisfying p *at any next state*.
- x is *a set variable in* ***recursive*** *formulae*.
- $\mu x.\, \psi' \vee \Box x \equiv \psi' \vee \Box\psi' \vee \Box(\psi' \vee \Box\psi') \vee \cdots$
  ➡ set of states satisfying ψ' or *at any next state*, or **recursively**, *at any next state*, or ...
- $\nu x.\, (\psi \to (\cdots)) \wedge \Box x$
  $\equiv (\psi \to (\cdots)) \wedge \Box(\psi \to (\cdots)) \wedge \Box\big((\psi \to (\cdots)) \wedge \Box(\psi \to (\cdots))\big) \wedge \cdots$
  ➡ set of states satisfying $(\psi \to (\cdots))$ and *at any next state*, and **recursively**, *at any next state*, and ...

Modal μ-calculus can be used to express many properties. For example,

(1) $\nu x.\, (\psi \to (\mu x.\, \psi' \vee \Box x)) \wedge \Box x$.

Each component represents:

- $\mu x.\, \psi' \vee \Box x$ means ψ' is satisfied *at a future point on any path*.
- $\nu x.\, (\psi \to (\cdots)) \wedge \Box x$ means if ψ is satisfied then $(\cdots)$ is *always* satisfied *on any path*.

Hence, (1) means that, **for any case, if ψ is satisfied, then ψ' is satisfied sometime later**.

NOTICE. Some properties *cannot be expressed in a subset of modal μ-calculus* known as CTL or LTL. However, our algorithm works as efficiently as that for CTL or LTL if the specification can be also written in CTL or LTL.

Abstractly interpreting an RNN

is done by *tracing all states with possible input vectors*:

- Calculate the set of states not satisfying the specification.
- It is represented with a *constrained zonotope closed under* addition, matrix application, *solving a linear equation*, intersection, and *bounded monotone element-wise activation functions*.

A counterexample is generated

with backpropagation: *calculating each input vector to force each state vector to be contained in the set computed in step (A)*. Those inputs *cause an RNN to not satisfy the specification*; hence, it is a possible attack pattern.

NOTICE. Like an RNN having a continuous state space, it is impossible to enumerate any path as a counterexample if it exists, so only formulae not containing $\Diamond$ are handled.