Todo list

Towards Reinforcement Learning of Invariants for Model Checking of Interlockings

Ben Lloyd-Roberts, Phillip James, Michael Edwards
Department of Computer Science, Swansea University UK
Email: {ben.lloyd-roberts, p.d.james, michael.edwards}@swansea.ac.uk

***Abstract*—The application of formal methods to verify that railway signalling systems operate correctly is well established within academia and is beginning to see real applications. However, there is yet to be a substantial impact within industry due to current approaches often producing false positive results that require lengthy manual analysis. It is accepted that invariants, properties which hold for all states under which a signalling system operates, can help reduce occurences of false positives. However automated deduction of these invariant remains a challenge. In this work we report on using reinforcement learning to explore state spaces of signalling systems and generate a dataset of state spaces from which we envisage invariants could be mined. Our results suggest the viability of reinforcement learning in both maximising state space coverage and estimating the longest loop free path for state spaces. Proximal Policy Optimisation (PPO) demonstrates the most stable learning, particularly in large environments where the optimal behaviour function is most complex. Whereas, distributed, multi-agent algorithms, such as Asynchronous Advantage Actor-Critic (A3C) result in the greatest state coverage.**

1. Introduction

Model checking is a formal verification technique stemming from the need to systematically check whether certain properties hold for different configurations (states) of a given system. Given a transition system T and a formula (or property) F, model checking attempts to verify through refutation that $s \vdash F$ for every system state $s \in T$, such that $T \vdash F$.

The application of model checking in order to verify railway interlockings is well established within academia and is beginning to see real applications in industry. As early as 1995, Groote et al. [1] applied formal methods to verify an interlocking for controlling the Hoorn-Kersenbooger railway station. Newer approaches to interlocking verification have also been proposed in recent years [2], [3], [4]. This includes work by Linh et al. [5] which explores the verification of interlockings written in a similar language to Ladder Logic using SAT-based model checking. In spite of this, such approaches still lack widespread use within the Rail industry.

In particular, one of the limitations of such model checking solutions is that verification can fail due to over approximation, typically when using techniques such as inductive verification [6]. Inductive verification fails to consider whether system states which violate a given property are indeed reachable by the system from a defined initial configuration. These false positive often require manual inspection. One solution is to introduce so-called invariants to suppress false positives [7]. Invariants are properties that hold for subregions of the state space. The aim is to introduce invariants that help bound the region of reachable states when model checking. However generating sufficiently strong invariants automatically is complex [8].

In this work we take first steps towards using machine learning to generate invariants by providing a first formal mapping of interlocking based state spaces to a reinforcement learning (RL) environment. We then explore how such state spaces can be generated in a controlled manner to test the scalability of our approach on environments where the number of reachable states is known. Finally we provide an analysis of how various reinforcement learning algorithms can be used to effectively explore state spaces in terms of their coverage. We see this as a first step towards mining invariants from such state spaces as this approach would indeed require reasonable coverage. Finally we reflect upon future works in directing our approach to improve exploration and learn invariants from a dataset of state sequences generated by our RL agents.

2. Preliminaries

We now briefly discuss model checking of railway interlockings and reinforcement learning. For further details we refer the reader to [9], [10] and [11], [12], [13] respectively.

2.1. Ladder Logic & Interlockings

Interlockings serve as a filter or 'safety layer' between inputs from railway operators, such as route setting requests, ensuring proposed changes to the current railway state avoid safety conflicts. As a vital part of any railway signalling system, interlockings are critical systems regarded with the highest safety integrity level (SIL4) according to the CENELEC 50128 standard.

Ladder logic is a graphical language widely used to program Programmable Logic Controllers [**?**] and in particular the Siemens interlocking systems we consider in this work. From an abstract perspective, ladder logic diagrams can be represented as propositional formulae. Here we follow the definition of James et al [10]. A ladder logic rung consists

of the following entities. *Coils* represent boolean values that are stored for later use as output variables from the program. A coil is always the right most entity of the rung and its value is computed by executing the rung from left to right. *Contacts* are the boolean inputs of a rung, with *open* and *closed* contacts representing the values of un-negated and negated variables respectively. The value of a coil is calculated when a rung fires, making use of the current set of inputs – input variables, previous output variables, and output variables already computed for this cycle – following the given connections. A horizontal connection between contacts represents logical conjunction and a vertical connection represents logical disjunction.

A interlocking executes such a program from top-to-bottom over and over, indefinitely.

More formally, following [10] a ladder logic program is constructed in terms of disjoint finite sets I and C of input and output/state variables. We define $C' = \{c' \mid c \in C\}$ to be a set of new variables in order to denote the output variables computed by the interlocking in the current cycle.

Defn 1. Ladder Logic Formulae: A ladder logic formula ψ is a propositional formula of the form

$$\psi \equiv ((c'_1 \leftrightarrow \psi_1) \wedge (c'_2 \leftrightarrow \psi_2) \wedge \ldots \wedge (c'_n \leftrightarrow \psi_n)$$

Where each conjunct represents a rung of the ladder, such that the following holds for all $i, j \in \{1, \ldots, n\}$:

- $c'_i \in C'$ (i.e. c' is a coil)
- $i \neq j \rightarrow c'_i \neq c'_j$ (i.e. coils are unique)
- $vars(\psi_i) \subseteq I \cup \{c'_1, \ldots, c'_{i-1}\} \cup \{c_i, \ldots, c_n\}$ (i.e. the output variable c'_i of each rung ψ_i, may depend on $\{c_i, \ldots, c_n\}$ from the previous cycle, but not on c_j with $j < i$, due to the nature of the ladder logic implementation, those values are overridden.)

Introduce ladder program

(CROSSING′ ↔ (REQ ∧ ¬ CROSSING),
REQ′ ↔ (PRESSED ∧¬ REQ),
TL_1_G′ ↔ ((¬ CROSSING′) ∧ (¬ PRESSED ∨ REQ′)),
TL_2_G′ ↔ ((¬ CROSSING′) ∧ (¬ PRESSED ∨ REQ′)),
TL_1_R′ ↔ CROSSING′,
TL_2_R′ ↔ CROSSING′,
PL_1_G′ ↔ CROSSING′,
PL_2_G′ ↔ CROSSING′,
PL_1_R′ ↔ ¬ CROSSING′,
PL_2_R′ ↔ ¬ CROSSING′,
AUDIO′ ↔ CROSSING′)

2.2. Transition Systems and Model Checking for Ladder Logic

For this work, we have concentrated on trying to produce invariants for the approaches taken by Kanso et al. [9] and James et al. [10]. Here we include their model of ladder logic based railway interlocking programs as we use this as a basis for defining a learning environment.

Building upon the propositional representation of a ladder logic program given in Section 2.1, we can define, following [10], the semantics of a ladder logic program in terms of labelled transition systems.

Let $\{0, 1\}$ represent the set of boolean values and let

$$Val_I = \{\mu_I \mid \mu_I : I \rightarrow \{0, 1\}\} = \{0, 1\}^I$$
$$Val_C = \{\mu_C \mid \mu_C : C \rightarrow \{0, 1\}\} = \{0, 1\}^C$$

be the sets of valuations for input and output variables.

The semantics of a ladder logic formula ψ is a function that takes the two current valuations and returns a new valuation for output variables.

$$[\psi] : Val_I \times Val_C \rightarrow Val_C$$
$$[\psi](\mu_I, \mu_C) = \mu'_C$$

where μ'_C is computed as follows: the value of each variable c_i is computed using the ith rung of the ladder, ψ_i, using the valuations μ_C and μ_I from the last cycle and the current valuations restricted to those evaluated be fore the current variable. We refer the reader to [10] for full details.

Defn 2. Ladder Logic Labelled Transition System: We define the labelled transition system $LTS(\psi)$ for a ladder logic formula ψ as the tuple $(Val_C, Val_I, \rightarrow, Val_0)$ where

- $Val_C = \{\mu_C | \mu_C : C \rightarrow \{0, 1\}\}$ is a set of states.
- $Val_I = \{\mu_I | \mu_I : I \rightarrow \{0, 1\}\}$ is a set of transition labels.
- $\rightarrow \subseteq Val_C \times Val_I \times Val_c$ is a labelled transition relation, where $\mu_C \xrightarrow{\mu_I} \mu'_C$ iff $[\psi](\mu_I, \mu_C) = \mu'_C$.
- $Val_0 \subseteq Val_C$ is the set of initial states.

We write $s \xrightarrow{t} s'$ for $(s, t, s') \in R$. A state s is called *reachable* if $s_0 \xrightarrow{t_0} s_1 \xrightarrow{t_1} \ldots \xrightarrow{t_{n-1}} s_n$, for some states $s_0, \ldots, s_n \in Val_C$, and labels $t_0, \ldots, t_{n-1} \in Val_I$ such that $s_0 \in Val_0$.

Consider Figure **??**, which, within one Figure, illustrates multiple models of a simple ladder logic program for controlling a pelican crossing[1]

Remove footnote if we're keeping the ladder code in sect 2.1

as considered by James et al [10]. One model highlighted is, as defined, a ladder logic LTS. We can see that states contain Boolean valuations for the ladder logic variables (for the LTS model we note the input variables below the dashed line at the bottom of each state are not included in the state variables). For example state S1 shows that the variable CROSSING is 0 (i.e. false) in that state. Transitions are labelled with (for our purposes here the blue labels) inputs and their Boolean valuations. For instance the arrow from S1 to S2 is labelled with the input PRESSED= 1. Finally we can also see one initial state, state S0 (the state with dotted edges), where all variables are set to false.

2.3. Reinforcement Learning and MDPs

Reinforcement Learning (RL) is a machine learning paradigm with demonstrably impressive capacity for modelling sequential decision making problems as the optimal

1. Here we omit the ladder logic code and refer the reader to [10]

control of some incompletely-known Markov Decision Process (MDP) [14].

Defn 3. Markov Decision Process A finite discounted Markov Decision Process M is a five tuple $(S, \mathcal{A}, P_a(s, s'), R_a(s, s'), \gamma)$, where

- S, is a finite set of states, known as the observation space or state space, representing the model state at discrete time steps.
- $\mathcal{A}$, describes the action space; a set of actions performable at discrete time steps, used to compute new states from the observation space.
- $P_a(s, s') = Pr(s_{t+1} = s' | s_t, a_t)$, describe state transition probabilities; the likelihood of observing state s_{t+1} given action a_t is taken from state s_t.
- $R_a(s, s')$ is a reward function feeding a scalar signal, r back to the agent at each time step t.
- $\gamma \in [0, 1]$ is a discount scalar successively applied at each time step.

In an RL setting we refer to the MDP as our environment, $\mathcal{E}$ where through simulation, software agents sample actions $a \in \mathcal{A}$, observe changes in states, $s \in S$ and learn to optimise some objective based on rewards, r issued over discrete time steps t. Simulation can be continuous, where agents indefinitely interact with the environment until some termination criterion is met, such as reaching some reward threshold. Alternatively tasks may be episodic, where training is conducted over a sequence of episodes, defined by a finite number of time steps. It is implicitly assumed ensuing algorithmic descriptions or problem formulation in this work refers to episodic cases. An agent's trajectory, $\tau = (s_1, a_1, r_1, s_2, a_2, r_2, ..., s_h, a_h, r_h)$, summarises experience accumulated over a single episode or continuous training run. Here, h refers to the horizon; a time step beyond which rewards are no longer considered. Rewards observed from time step t up to some terminal time step T, are denoted $G_t = \sum_{i=t}^{T} \gamma^{i-t} r_i$, and referred to as the return. Discount factor $\gamma \in [0, 1]$ downscales rewards over return time steps while future rewards are reduced as the discount exponent $i - t$ increases. This helps prioritise immediate rewards over distant ones and enumerate returns over a potentially infinite horizon.

Two principle challenges of RL are those of prediction and control. The first refers to approximating the value function used to estimate state values $v(s)$ or state-action values $q(s, a)$, describing the benefit of being in that state. Values are differentiable and updated based on an empirical average, known as the expectation, taken over observed returns, $\mathbb{E}[G_t | S_t = s, A_t = a]$. In other words the expectation over return G_t depends being in state $S_t = s$, having taken action $A_t = a$ during the present time step. Observed returns depend on the action(s) sampled at discrete time steps. The second challenge of control concerns optimising this selection process via the policy $\pi(a|s)$; a probability distribution mapping states to actions most likely to maximise the reward objective. Ultimately an optimal value function will converge to an optimal policy [15]. In practice to scale with complex environments these functions are parametric and approximated with gradient-based methods, such as stochastic gradient ascent.

Determining the reachable state space for ladder programs is often intractable, making the complete MDP model unknown to us without computing transitions for all state-action pairs. Consequently we trial several gradient-based learning algorithms using deep neural network state representations to approximate both value function and policy subject to the reward objective. First a simple Deep Q-Network [11] is trained on small environments before applying more advanced approaches. Among this family of approximate reinforcement learning algorithms are policy gradient methods [16]. Such algorithms utilise parametrised policies for control, where $\pi(a|s, \theta) = Pr(A_t = a | S_t = s, \theta_t = \theta)$ describes action selection, without need of value function estimates. Actor-critic methods also approximate a parametric, typically state-value, function $\hat{v}(s, \omega)$ separating learning of action selection from predicting value estimates. Gradient updates are performed with respect to parameter vector θ and objective function $J(\theta)$. Definitions of $J(\theta)$ vary depending on the learning algorithm. In this work we explore the efficacy of three principle actor-critic algorithms; Proximal Policy Optimisation (PPO) [12], Advantage Actor-Critic (A2C) and its asynchronous counterpart (A3C) [13] in navigating our set of generated LLPs. We discuss the merits and drawbacks of each approach for our setting in Section 5.1.

3. Related Work

Here we briefly highlight key contributions within related literature, addressing the invariant finding problem for interlocking programs and contemporary RL strategies for environment exploration.

3.1. Invariant Finding

Elaborate on these citations

From software engineering techniques [17], [18] to hybrid methods incorporating machine learning [19], researchers have proposed various approaches to invariant finding with varying degrees of success. IC3 [20] is one of the most successful approaches for model checking with invariants. IC3 makes use of of relatively simple SAT problems to incrementally constrain a state space towards only reachable states. In this scenario, IC3 operates only at the Boolean level of the abstract state space, discovering inductive clauses over the abstraction predicates. It has been applied to verification of software [21] and indeed, in the context of hardware model checking [22]. Although we note that this approach contains the state-space relative to a given property for verification. In our work, we aim to explore invariants that can be mined independent of the given property.

3.2. Exploration in Reinforcement Learning

Maintaining the desired balance between exploring unobserved state-action pairs for information maximisation and exploiting knowledge of the environment to further improve performance remains a challenge in RL research. Means of improving state exploration has seen particular interest in software or user testing communities. In [23], Bergdahl et al. apply vanilla PPO in an episodic 3D environment, incentivising state space coverage to automatically identify in-game bugs and exploits. Recently, Cao et al. [24] used a curiosity function based on an empirical count of state-action pairs to maximise traversal of state transitions for specific human-machine interfaces, using vanilla Q-learning with an ϵ-greedy exploration. Similar applications in web application testing [25] simulate user actions to navigate site structures, recalling which state-action pairs the model is most 'uncertain'. State transitions with the greatest uncertainty are then prioritised for exploration when backtracking from state loops. Again, authors use vanilla Q-learning and count-based reward scaling during Q-function updates [26]. Authors of [27] decompose exploration tasks over large, adaptive and partially observable environments into two sub-problems; adaptive exploration policy for region selection and separate policy for exploitation of an area of interest. Other works [28] incorporate recurrent networks [29] in policy design, using temporality to recall the performance of past actions and their subsequent consequences according to the reward function. In robotics research [30], Apuroop et al. apply PPO to control hTrihex robot navigation in procedurally generated environments, achieving near human level performance. Novel training algorithms [31] have also been devised to learn combinatorial algorithms over large graph structures comprising billions of nodes. We incorporate trends in the literature such as state-of-the-art learning algorithms sporting the best performance in research tasks and using environment reset logic to discourage early convergence. We detail this approach further in Section 5.

4. Mapping Formal Methods to Reinforcement Learning

Before any attempts toward invariant finding can be made, it is essential our model in the reinforcement learning setting captures the structure of model checking on ladder programs. In this section we introduce a faithful mapping that, given a ladder logic LTS, constructs an MDP model where reinforcement learning can be applied. Similarly, any invariant finding approach will require agents that maximise state space coverage. To explore this, we also introduce an approach for generation of state spaces with known metrics such as number of states and state space depth.

4.1. Ladder Logic Markov Decision Process

We now define the finite Markov Decision Process (MDP), or environment $\mathcal{E}$ used to represent the LLP. *Defn 4. Ladder Logic Markov Decision Process* A Ladder Logic MDP $M(\psi) = (S, \mathcal{A}, P_a(s, s'), R_a(s, s'), \gamma)$ is a five tuple, where our observation space is the union of program inputs and ladder variables, and:

- $S = Val_C \cup Val_I$.
- $\mathcal{A} = Val_I$.
- $P_a(s, s') = Pr(s_{t+1} = s' | s_t, a_t)$
- $R_a(s, s')$ is a reward function feeding a scalar signal back to the agent at each time step t.
- $\gamma \in [0, 1]$ is a discount scalar successively applied at each time step.

Here we note that any unique valuation of Val_C under the dynamics of a LLP constitutes a distinct state. Our action space, describes the set of ladder logic inputs used to compute new valuations after program execution. $P_a(s, s') = Pr(s_{t+1} = s' | s_t, a_t)$, describes the state transition function in terms of probabilities of observing s_{t+1} given action a_t is taken from state s_t. As here, more transitions are available than described by the ladder logic program, we use this probability distribution to ensure transitions match those defined by the ladder logic (essentially this will be 1 for transitions dictated by the ladder logic program and 0 for transitions that are not).

Subsequently as agents build a policy $\pi(s|a, \theta)$ according to $R_a(s, s')$ and state transitions observed under $P_a(s, s')$, the environment unfolds as a set of reachable states that mirror those of the ladder logic LTS.

Considering Figure 1, we observe differences in how agent action selection is represented compared to LTS transition labels. Where PRESSED and ACT_1 are ladder logic inputs, the indices of an agents action space refer to selecting one of the following valuations: [PRESSED=True, PRESSED=False, ACT_1=True, ACT_1=False]. One index may evaluate to True (1) while the remainder are False (0). Following Figure 1, an agent starting its training episode from initial state $S0$ has four available actions (illustrated in red) and three states reachable, $S1, S2, S4$ within the next time step. Selecting action $[0, 1, 0, 0]$ denotes setting PRESSED=False, observing a 'new' state $S1$, and receiving positive reward $+1$, completing the time step. For ladder logic MDP with N actions (LTS transition labels), there are at most $2N$ unique transitions (s, a, s'), thus the size of the action space from all states $s \in S$, is also $2N$.

Finally, our reward function and γ can be tuned to modify the learning objective. Aiming to maximise state space coverage we implement a reward scheme which positively rewards novel observations over distinct episodes, deterring loop traversal through episode termination and negative rewards. Consider the example trajectory τ = (S0, [0,1,0,0], +1, S1, [0,0,1,0], +1, S4, [0,1,0,0], +1, S5, [0,1,0,0], -1, S5). Computing the expected return on the next episode, starting from $S0$ and using observed rewards from the latest trajectory, discounting factor $\gamma = 0.99$ applies accordingly; $G_t = 1 + 0.99(1) + 0.99^2(1) = 1 + 0.99 + 0.9801$. In Section 5 we discuss these point further.

4.2. Environment Generation

Given exhaustive search of large state spaces is often computationally intractable, we have generated a set of ladder programs where the number of reachable states and recurrence reachability diameter are known. This enables us to analyse the performance of our approach against well understood state spaces. Using existing models of ladder logic structures as a base template [10], we derive progressively larger programs by sequentially introducing additional rungs. This way a constrained yet predictable pattern of growth is devised. If $|S(\psi_i)|$ represents the number of reachable states for a program ψ_i, a subsequently generated program ψ_{i+1} with one additional rung, has $2|S(\psi_i)| + 1$ reachable states. Through a series of training runs on each environment we record the number of states observed by workers to gauge the overall state space coverage. The following algorithm modifies the body of a pelican crossing ladder program referenced in [10]. For every i^{th} rung introduced to lengthen the base program, we define two additional variables, VAR_i and ACT_i. This effectively doubles the existing state space, increases the size of each state observation and introduces as increments the size of $\mathcal{A}$ by 2.

procedure GENERATE LADDER(n_rungs, $prog$)
  cond ← (¬ PRESSED ∧ ¬CROSSING) ∧ ¬ REQ)
  rung ← ACT_1 ∧ cond
  coil ← VAR_1 ↔ rung
  $i \leftarrow 1$
  while $i \leq$ n_rungs **do**
    $i \leftarrow i + 1$
    new_rung ← ACT_[i] ∧ (rung)
    new_coil ← VAR_[i] ↔ (new_rung)
    append new_coil **to** $prog$
  end while
end procedure

Applying the above algorithm with $n\_rungs = 0$ produces the complete state space of Figure 1. For readability and illustration we have obfuscated three states from the original ladder program in Section 2.1, abstracted as the 'Pelican State Space'. Similarly, five states and their connected transitions are abstracted as the 'Extended State Space'.

Not sure if the following is already explained

Again, dashed edges between states denote transitions invoked by actions we have introduced while hard lined edges are invoked by the valuations of PRESSED. Transition labels from the LTS are highlighted in blue, while action selection according to the RL agent is shown in red. Results presented in Section 5 are based on a set of generated programs ranging from 2^{14} to 2^{50} states, referenced in Table 1.

5. Results

We now present a set of results from applying our approach to a series of generated ladder programs, modelled as learning environments. The following section is divided into three parts, first discussing the merits and drawbacks of the respective learning algorithms. Second, we present results from applying the multi agent A3C algorithm to our full set of generated ladder programs, shown in Table 1. Finally we discuss performance differences between single agent implementations of A2C and PPO, as they exhibited divergent learning objectives.

5.1. Algorithmic Trends

Make this coherent

Initially to test our model we train a simple DQN agent on a set of generated programs, up to 2^{20} theoretic states and 127 reachable. For implementation details regarding The agent often observes complete state coverage within the time taken for the more advanced A2C and PPO algorithms to determine state space depth. DQNs exhibited difficulty scaling well to state space depth learning objective without significant hyperparameter tuning, failing to converge performance on the smallest environments. It is likely using DQN without a prioritised replay buffer [32] leads to poor training sample efficiency when applying network updates. We expect agents exploring ladder program states to rarely observe the greatest depth, meaning episodes resulting in optimal rewards may never be sampled for training. Bootstrapping the action-value function via a target network also introduces an additional challenge in selecting an update interval frequent enough to avoid overestimating state-action values. Similarly, the use of replay memory requires an initial training delay to populate the buffer before Q-Network updates. Setting the gradient step size is likely dependent on the learnable function complexity, which we'd expect to increase with environment size. The off-policy nature of DQNs constrains network updates with trajectories produced under other, presumably less optimal, policies. Governing the DQN exploration rate explicitly using an ϵ-greedy strategy seems unstable given this could reduce random action exploration with large subregions of an environment unobserved. While DQN demonstrates the potential of RL algorithms in learning policies for prediction and control, they are unlikely to scale sufficiently to interlocking state spaces for the purposes of invariant finding.

Two Actor critic methods, single agent Advantage actor-critic (A2C) and its asynchronous multi-agent counterpart (A3C) are applied to all 19 generated environments. Having several workers and an on-policy learning algorithm removes need of replay memory and any training delay to accumulate sufficient experience. Actor and critic networks, approximating the behaviour policy and value function, provide better convergence guarantees at the cost of some additional complexity in learned parameters. Paired with randomised reset logic both algorithms accumulate experience faster than conventional DQN, achieving good coverage on a range of medium to large state spaces. Actor-critic algorithms, while improved by more informed value estimation and policy iteration, they are also susceptible to performance collapse in sufficiently complex environments

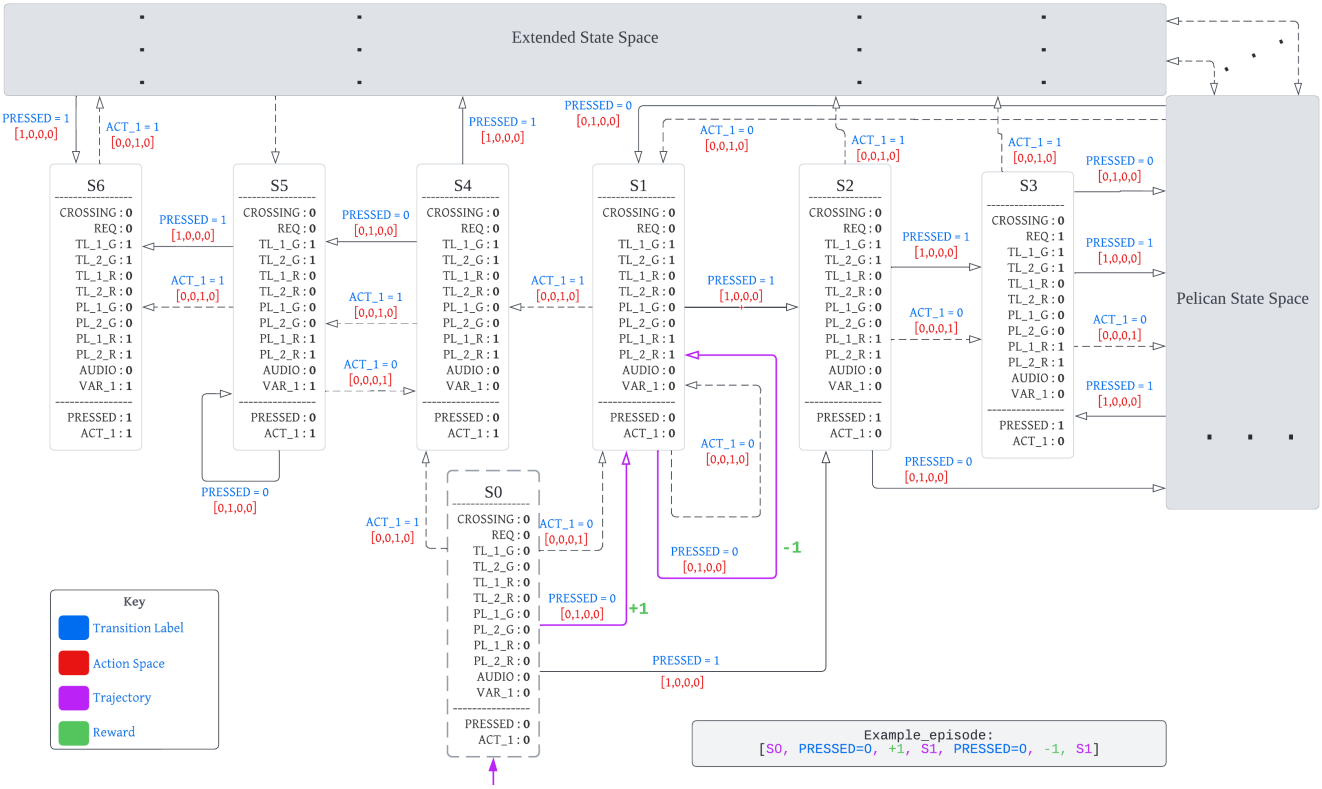


Figure 1: Simplified state space representation of generated ladder logic program with one additional input variable ACT_1 and one output coil VAR_1.

or over long training periods. Parameter updates have a tendency to push the policy into an unfamiliar region of policy space from which subsequent updates are unable to recover, potentially 'collapsing' the model. Unsurprisingly A3C outperforms the single agent variant and sports the best coverage metric among all policy gradient methods applied. A3C and its results are discussed in depth in Section 5.2.

It is for this reason we move to Proximal Policy Optimisation (PPO), to constrain the magnitude of gradient steps during the parameter update. It is our expectation that limiting policy, and potentially value function, updates within a clip range [12] or according to KL-divergence between the current and most recent policy [33], making the model more resilient to collapse. Trust region methods generally require fewer hyperparameter adjustments being more stable learning strategies. We observe this in Section 5.3, where PPO demonstrates slow but linear state exploration while appearing to maximise state depth within a known subregion.

TABLE 1: A3C Coverage Metrics

Environment			Agent	
States (Theoretical)	States (Reachable)	Actions	Depth	Coverage
2^{14}	15	2	14	100.0
2^{16}	31	4	28	100.0
2^{18}	63	6	48	100.0
2^{20}	127	8	33	100.0
2^{22}	255	10	76	100.0
2^{24}	511	12	49	100.0
2^{26}	1023	14	306	100.0
2^{28}	2047	16	538	100.0
2^{30}	4095	18	1418	99.731
2^{32}	8191	20	1712	96.532
2^{34}	16383	22	1498	95.550
2^{36}	32767	24	2879	84.694
2^{38}	65535	26	1969	89.071
2^{40}	131071	28	2692	82.884
2^{42}	262143	30	1406	76.033
2^{44}	524287	32	1782	62.137
2^{46}	1048575	34	1593	64.053
2^{48}	2097151	36	1598	57.547
2^{50}	4194303	38	2566	41.483

5.2. Asynchronous Exploration

Preliminary results applying the A3C algorithm to a number of generated programs are outlined in Table 1. Training was distributed among 32 CPU worker threads. 'Actions' referenced in the third column refer to the number of possible assignments over input variables in each ladder program, from every state. Depth, the first agent column, refers to the greatest number of steps taken before repeating observations in the environment, across all workers.

Coverage metrics are expectedly maximised for environments with a small number of reachable states with

acceptable levels of coverage for programs with a theoretical size up to 2^{40}. Interestingly, we observed longer training durations without early stopping occasionally increased coverage beyond a certain threshold. It is possible workers learn an optimal search strategy within a subregion of the state space. Additionally, performance in terms of max depth and states reached increased by approx. 5% when decreasing the total number of episodes from 3×10^5 to 1.5×10^5 episodes. This may be a product of random episode initialisation spawning workers in more desirable states where stochastic action sampling happened to lead to unfamiliar subregions of the environment.

Performance in terms of cumulative reward which failed to maximise coverage often increased linearly before collapsing to some suboptimal reward. This may be due to tendencies for large network updates to shift the network gradients into a bad local minima, from which performance does not recover within the allotted training duration. The on-policy nature of actor critic means trajectories generated via an old policy are no longer sampled during minibatch updates for the current policy, thus biasing behaviour to the most recent model updates and introducing sample inefficiency. Adding on policy memory strategies [34] may help avoid this in future applications

Given the A3C algorithm requires workers to asynchronously update their shared network every $T_{\max}$ steps or on episode termination, larger values for $T_{\max}$ consolidate more information regarding worker trajectories before applying gradient updates to their local network. We found the most significant improvements to performance in terms of coverage metrics and increasing the k bound when introducing workers to larger environments, was lower update frequencies and random start state initialisation. Prior to these adjustments workers, irrespective of their number, seldom covered 80% of most smaller environments. Similarly, for the largest environment with 2^{50} states, coverage improved from 3.2% to 41.48%

5.3. Trust Regions and Reward Shaping

Advantage actor critic methods seem to perform better in terms of coverage, particularly distributed variants. The multi agent nature paired with no update clipping may allow for greater exploration at the cost of training stability. Having observed the best coverage performance from A3C, we select a number of ladder programs to compare single agent algorithms, namely A2C and PPO, in terms of their training behaviour and evaluation performance. Having designed our reward function as one of continued state novelty we expected PPO to outperform A2C, covering a greater proportion of states after A2C performance either plateaus or collapses.

Figure 2(a) shows the number of reachable state discovered by the agent during the first 5 hours of training on a ladder program with 19 additional rungs. A2C increases state coverage significantly faster than PPO but considering Figure 2(b), struggles to learn which transitions reveal the max state depth within its discovered subregion.

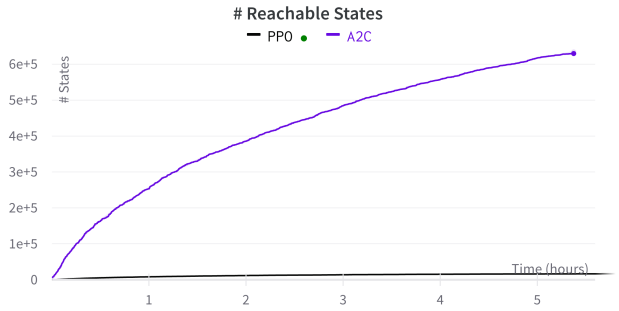
Again, Figure 2(b) demonstrates PPO optimises better to the actual reward objective, maximising exploration of state-action pairs for the explored subregion while slowly increasing state coverage. The surrogate clipped objective used during the gradient update may contribute to how agents introduce new states by steadily exploring uncertain state-action pairs. Figure 2(c) may also suggest this, in that we observe the mean entropy loss for PPO is significantly more stable than that of A2C, indicating the policy under PPO maintains higher levels of uncertainty. Conversely, A2C mean entropy loss appears to converge toward 0, indicating its prediction are more certain despite very poor performance in terms of the actual reward objective. Similarly the value function loss is more stable for PPO compared to A2C. This measures the TD error between the current value function and actual observed returns.

Model performance across all learning algorithms appears sensitive to network update frequency and experience accumulation during rollout, prior to each update. This is certainly a property to be conscious of as it may translate interlocking programs, making this parameter dependent on the state space structure, where the total number of reachable states or longest acyclic path is unknown to us. In future we would like to trial this set of algorithms, as well as some hybrid approaches [35], using a modified reward function based on global state coverage or uncertainty maximisation.

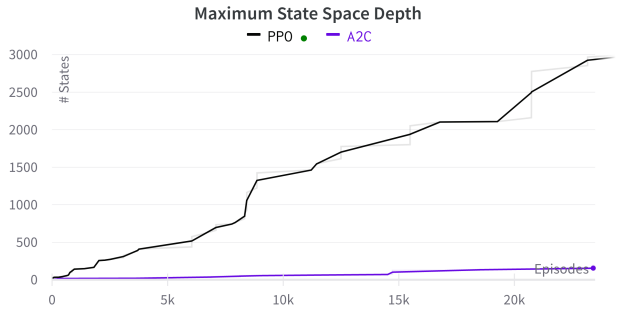
6. Conclusion & Future Work

In this work we have taken first steps towards using machine learning to generate invariants by providing a first formal mapping of interlocking based state spaces to a reinforcement learning (RL) environment. In addition, we have applied asynchronous and trust region deep reinforcement learning methods to programtically generated state spaces and analysed their ability in terms of state coverage and state space depth. Our findings highlight that RL approaches can be succesfully rewarded to explore a large percentage of a given state space in terms of state coverage, however that incentivising depth based exploration is more challenging. As any machine learning approach to finding invariants will likely need to explore such a state space these results show the credability of such an appraoch. In our subsequent works we envisage the current learning framework to serve as a means of dataset generation, from which patterns or sequences of can be mined from agent trajectories.

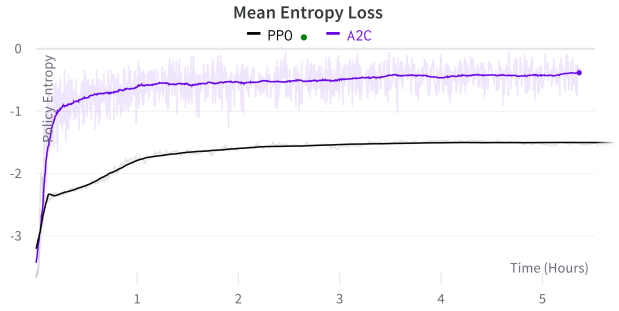
In light of our findings, we also aim to improve several aspects of our approach, predominantly concerning learning stability, sample efficiency and training speed. The low dimensionality of our state space representation may allow us to introduce count-based exploration models to dampen the reward issued for repeated observations [36]. Intrinsic motivation has also demonstrated success in shaping rewards to boost environment exploration [37]. Such further tuning of our reward scheme may also lead to improved coverage performance using PPO, where the learning objective may be less complex than one pursuing state space depth.



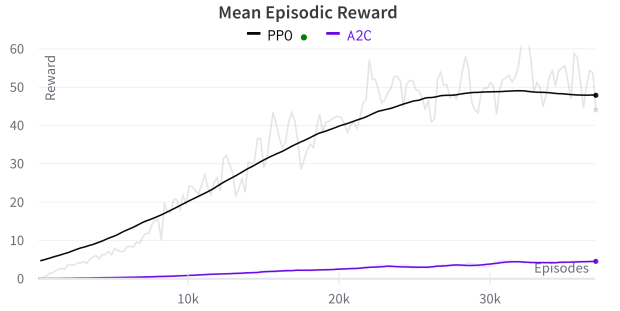
(a) Number of reachable states discovered during five hours of training.



(b) Max state space depth observed over 25 thousand training episodes.



(c) Mean policy entropy loss. The negative average entropy output by the policy network.



(d) Rolling average of reward for 35 thousand training episodes.

Figure 2: Training and evaluation metrics during application of PPO and A2C algorithms on generated state space of a theoretic size 2^{50}.

Applying modified algorithms such as IMPALA [38] could improve both the sample efficiency over our A3C implementation and increase model capacity for deeper network architectures and robustness against hyperparameter adjustments. The adoption of a Long Short-Term Memory model (LSTM) could potentially improve performance by predicting candidate state sequences at discrete time steps, adding some temporality to the learned features.

Acknowledgments

We thank Tom Werner and Andrew Lawrence at Siemens Mobility UK & EPSRC for their support in these works.

References

[1] J. F. Groote, S. F. van Vlijmen, and J. W. Koorn, "The safety guaranteeing system at station hoorn-kersenboogerd," in *COMPASS'95 Proceedings of the Tenth Annual Conference on Computer Assurance Systems Integrity, Software Safety and Process Security'*. IEEE, 1995, pp. 57–68.

[2] A. Fantechi, W. Fokkink, and A. Morzenti, "Some trends in formal methods applications to railway signaling," *Formal methods for industrial critical systems: A survey of applications*, pp. 61–84, 2012.

[3] A. Ferrari, G. Magnani, D. Grasso, and A. Fantechi, "Model checking interlocking control tables," in *FORMS/FORMAT 2010*. Springer, 2011, pp. 107–115.

[4] A. E. Haxthausen, M. L. Bliguet, and A. A. Kjær, "Modelling and verification of relay interlocking systems," in *Monterey Workshop*. Springer, 2008, pp. 141–153.

[5] L. H. Vu, A. E. Haxthausen, and J. Peleska, "Formal modeling and verification of interlocking systems featuring sequential release," in *International Workshop on Formal Techniques for Safety-Critical Systems*. Springer, 2014, pp. 223–238.

[6] H. Post, C. Sinz, A. Kaiser, and T. Gorges, "Reducing false positives by combining abstract interpretation and bounded model checking," in *2008 23rd IEEE/ACM International Conference on Automated Software Engineering*, 2008, pp. 188–197.

[7] M. Awedh and F. Somenzi, "Automatic invariant strengthening to prove properties in bounded model checking," in *2006 43rd ACM/IEEE Design Automation Conference*, 2006, pp. 1073–1076.

[8] G. Cabodi, S. Nocco, and S. Quer, "Strengthening model checking techniques with inductive invariants," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 1, pp. 154–158, 2009.

[9] K. Kanso, F. Moller, and A. Setzer, "Automated verification of signalling principles in railway interlocking systems," *Electronic Notes in Theoretical Computer Science*, vol. 250, no. 2, pp. 19–31, 2009.

[10] P. James, A. Lawrence, F. Moller, M. Roggenbach, M. Seisenberger, A. Setzer, K. Kanso, and S. Chadwick, "Verification of solid state interlocking programs," in *International Conference on Software Engineering and Formal Methods*. Springer, 2013, pp. 253–268.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[13] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," 2016.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[15] F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 664–671.

[16] S. M. Kakade, "A natural policy gradient," *Advances in neural information processing systems*, vol. 14, 2001.

[17] M. L. Case, A. Mishchenko, and R. K. Brayton, "Automated extraction of inductive invariants to aid model checking," in *Formal Methods in Computer Aided Design (FMCAD'07)*. IEEE, 2007, pp. 165–172.

[18] S. Bensalem, Y. Lakhnech, and H. Saidi, "Powerful techniques for the automatic generation of invariants," in *International Conference on Computer Aided Verification*. Springer, 1996, pp. 323–335.

[19] P. Garg, D. Neider, P. Madhusudan, and D. Roth, "Learning invariants using decision trees and implication counterexamples," *ACM Sigplan Notices*, vol. 51, no. 1, pp. 499–512, 2016.

[20] A. R. Bradley, "Sat-based model checking without unrolling," in *Verification, Model Checking, and Abstract Interpretation*, R. Jhala and D. Schmidt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 70–87.

[21] A. Cimatti and A. Griggio, "Software model checking via ic3," in *Computer Aided Verification*, P. Madhusudan and S. A. Seshia, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 277–293.

[22] A. R. Bradley and Z. Manna, "Checking safety by inductive generalization of counterexamples to induction," in *Formal Methods in Computer Aided Design (FMCAD'07)*, 2007, pp. 173–180.

[23] J. Bergdahl, C. Gordillo, K. Tollmar, and L. Gisslén, "Augmenting automated game testing with deep reinforcement learning," in *2020 IEEE Conference on Games (CoG)*, 2020, pp. 600–603.

[24] Y. Cao, Y. Zheng, S.-W. Lin, Y. Liu, Y. S. Teo, Y. Toh, and V. V. Adiga, "Automatic hmi structure exploration via curiosity-based reinforcement learning," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2021, pp. 1151–1155.

[25] Y. Zheng, Y. Liu, X. Xie, Y. Liu, L. Ma, J. Hao, and Y. Liu, "Automatic web testing using curiosity-driven reinforcement learning," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, 2021, pp. 423–435.

[26] H. Tang, R. Houthooft, D. Foote, A. Stooke, O. Xi Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel, "# exploration: A study of count-based exploration for deep reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

[27] A. Peake, J. McCalmon, Y. Zhang, D. Myers, S. Alqahtani, and P. Pauca, "Deep reinforcement learning for adaptive exploration of unknown environments," in *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2021, pp. 265–274.

[28] K. G. S. Apuroop, A. V. Le, M. R. Elara, and B. J. Sheu, "Reinforcement learning-based complete area coverage path planning for a modified htrihex robot," *Sensors*, vol. 21, no. 4, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1067

[29] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[30] D. I. Koutras, A. C. Kapoutsis, A. A. Amanatiadis, and E. B. Kosmatopoulos, "Marsexplorer: Exploration of unknown terrains via deep reinforcement learning and procedurally generated environments," *Electronics*, vol. 10, no. 22, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/22/2751

[31] S. Manchanda, A. Mittal, A. Dhawan, S. Medya, S. Ranu, and A. Singh, "Learning heuristics over large graphs via deep reinforcement learning," *arXiv preprint arXiv:1903.03332*, 2019.

[32] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba, "Hindsight experience replay," *Advances in neural information processing systems*, vol. 30, 2017.

[33] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," 2017.

[34] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," 2017.

[35] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[36] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos, "Count-based exploration with neural density models," 2017.

[37] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel, "Vime: Variational information maximizing exploration," 2017.

[38] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," 2018.