TacticZero: Learning to Prove Theorems from Scratch with Deep Reinforcement Learning

{ Minchao Wu, Michael Norrish, Christian Walder and Amir Dezfouli }
Australian National University & Data61, CSIRO

Objectives

- A precise formulation of the practice of interactive theorem proving (ITP) in terms of Markov decision process (MDP), which enables learning tactic selection as well as proof search strategies.
- An RL architecture using multiple recurrent and feed-forward neural network modules to solve the MDP. The architecture and learning algorithm we use are designed for handling large state and action spaces.
- Comparable performance to approaches that rely on examples from human experts available in the HOL4 system.

Introduction & Motivation

Existing approaches focus on learning from human example proofs and use fixed proof search strategies such as breadth first search. There are potential limitations with such approaches.

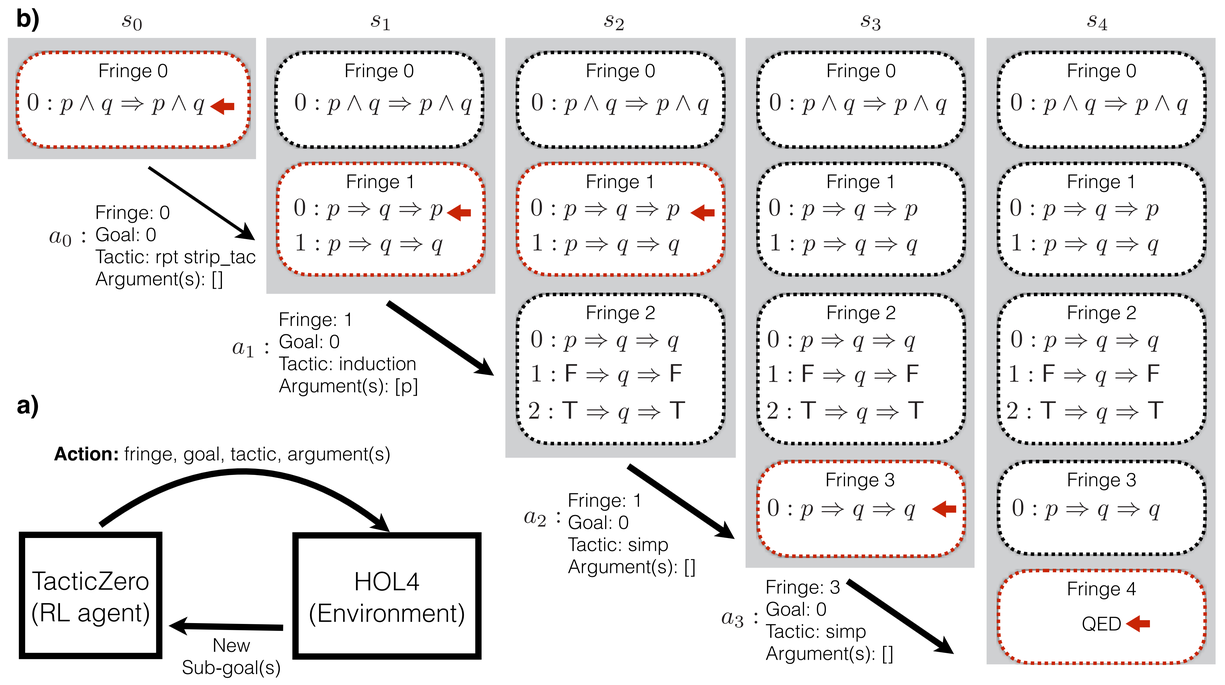
- Since human example proofs are used, the quality of the trained agent depends on the quality of human proofs. There are potentially infinitely many proofs of a theorem, but the agent is only able to learn one or two of them that exist in the library.
- Search strategies such as BFS are expensive in time and space, and are not quite human-like.

We want a learning framework that enables an agent to learn to **prove theorems without human examples**, and **learn proof search strategies by itself**. Moreover, such a framework should open the possibility of applying principled reinforcement learning algorithms to the problem of ITP.

References

[1] Laura Kovács and Andrei Voronkov. First-order theorem proving and Vampire. In Natasha Sharygina and Helmut Veith, editors, *Computer Aided Verification*, pages 1–35, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

MDP Formulation



Policy Structure

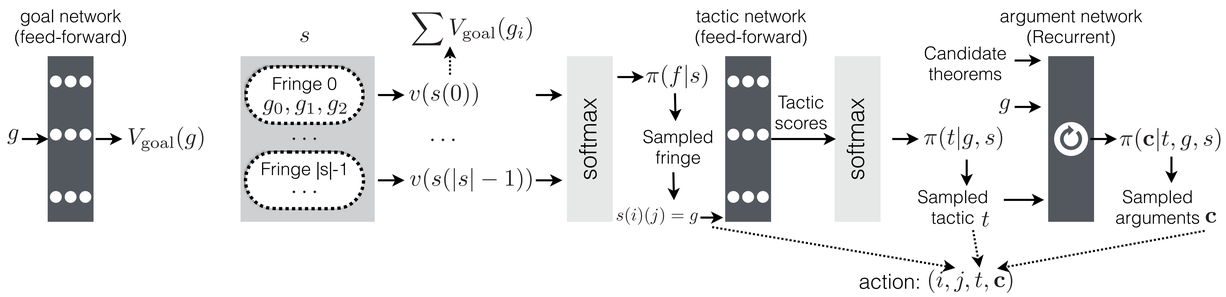


Figure 1: The policies are trained jointly using stochastic Monte Carlo policy gradient.

Argument Policy

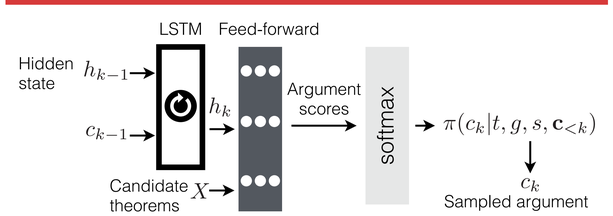


Figure 2: Recurrent selection of arguments.

Future Improvements

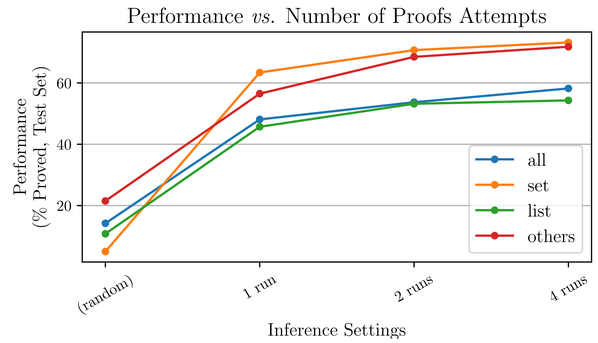
- Outsource the subtask of free generation of terms to a language model.
- Use a more principled off-policy algorithm such as IMPALA.
- Larger scale experiments by curriculum learning through the structured HOL4 library.

Proved Theorems

Table 1: Percentage of proved theorems by TacticZero trained on different datasets compared to that by the corresponding random rollouts and HOL(y)Hammer (with Vampire [1]) on unseen theorems.

Method	all	set	list	others
Random	14.2	5.0	10.8	21.5
TacticZero	62.3	**81.7**	61.7	**75.0**
Hammer	64.5	69.5	62.8	64.1

The performance of TacticZero increases when it is allowed to perform multiple proof attempts to a theorem. This is because some attempts might be unsuccessful due to stochastic policy.



Example Proof Search

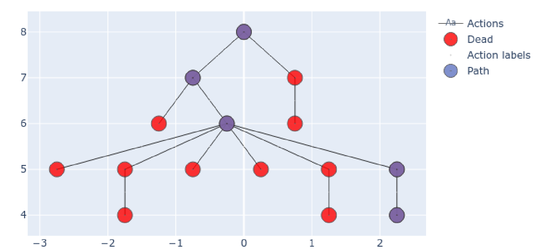


Figure 3: An example proof search (which is neither BFS nor DFS) of theorem $\forall\ x\ s.\ x \in s \Rightarrow \forall f.\ f(x) \in$ `IMAGE` $f\ s$. This particular proof was found in 13 steps. Red nodes represent the fringes that never lead to a successful proof, and blue nodes consist of a path from which a valid HOL4 proof can be re-constructed.