**Overview**

- **Motivation**
  - Deep reinforcement learning (RL) has been successfully applied to solve a number of challenging control tasks
  - However, it’s real-world applicability remains limited due to safety concerns in using learned, blackbox controllers

- **Our approach**
  - Decision tree controllers are easy to verify but hard to train
  - Use imitation learning to train a decision tree controller

**Background on Imitation Learning**

- **Naive algorithm**
  - **Step 1**: Use NN to generate states
  - **Step 2**: Use NN to label action for each state
  - **Step 3**: Use supervised learning to train DT

- **DAgger Algorithm (Ross 2011)**
  - **Problem**: DT makes mistakes and sees new states
  - **Solution**: Use NN to label states

- **Decision tree controllers are easy to verify but hard to train**

**Problem Formulation**

- **Input**
  - Markov decision process (MDP) $M = (S, A, T, R)$
  - Neural network (NN) controller $\pi_{NN}: S \rightarrow A$
  - $Q$ function, where $Q(s, a)$ measures how good action $a$ is in state $s$ (obtained from deep RL algorithms)

- **Output**
  - Decision tree (DT) controller $\pi_{DT}: S \rightarrow A$

**Evaluation**

- **Comparison to reinforcement learning for DTs (below, left)**
  - Fitted Q iteration (RL algorithm for learning decision trees)
  - Cart-pole control problem

- **Comparison to DAgger (below, right)**
  - Atari Pong (symbolic state space)

**Case Study: Verifying Toy Pong**

- **Toy pong**
  - **Problem**: $S = \mathbb{R}^5$, $A = \{\text{left, right, stay}\}$
  - **NN**: Trained using policy gradients, 600 neurons
  - **DT**: Extracted using VIPER, 31 nodes
  - **Correctness**: Never lets the ball leave the arena

- **Verification**
  - **Inductive invariant**: $s(0) \in \text{blue} \Rightarrow s(T) \in \text{blue}$ (below, left)
  - **Algorithm**: Dynamics and DT controller are piecewise linear, so we can encode correctness as an SMT formula

- **Results**
  - Solved by Z3 in < 5 seconds
  - Finds an error when ball starts on the right (below, right)
  - Fixed when paddle is slightly longer!

**References**

Ross, Gordon, & Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. AISTATS 2011.