Version of Record: https://www.sciencedirect.com/science/article/pii/S0893608022001009
Manuscript_cbf8c5b8df4a1d7871069f3a9a69e52c

# MoËT: Mixture of Expert Trees and its Application to Verifiable Reinforcement Learning

Marko Vasic[a], Andrija Petrovic[b], Kaiyuan Wang[c], Mladen Nikolic[d], Rishabh Singh[e], Sarfraz Khurshid[a]

[a] *The University of Texas at Austin, USA*
[b] *Singidunum University, Serbia*
[c] *Google, USA*
[d] *University of Belgrade, Serbia*
[e] *Google Brain, USA*

## Abstract

Rapid advancements in deep learning have led to many recent breakthroughs. While deep learning models achieve superior performance, often statistically better than humans, their adoption into safety-critical settings, such as healthcare or self-driving cars is hindered by their inability to provide safety guarantees or to expose the inner workings of the model in a human understandable form. We present MoËT, a novel model based on Mixture of Experts, consisting of decision tree experts and a generalized linear model gating function. Thanks to such gating function the model is more expressive than the standard decision tree. To support non-differentiable decision trees as experts, we formulate a novel training procedure. In addition, we introduce a hard thresholding version, MoËT$_{\text{h}}$, in which predictions are made solely by a single expert chosen via the gating function. Thanks to that property, MoËT$_{\text{h}}$ allows each prediction to be easily decomposed into a set of logical rules in a form which can be easily verified. While MoËT is a general use model, we illustrate its power in the reinforcement learning setting. By training MoËT models using an imitation learning procedure on deep RL agents we outperform the previous state-of-the-art technique based on decision trees while preserving the verifiability of the models. Moreover, we show that MoËT can also be used in real-world supervised problems on which it outperforms other verifiable machine learning models.

*Email address:* vasic@utexas.edu (Marko Vasic)

*Keywords:* Verification, Deep Learning, Reinforcement Learning, Mixture of Experts, Explainability

## 1. Introduction

Deep learning has achieved many recent breakthroughs, in challenging domains such as Go [1], and healthcare [2, 3] to name a few. Encoding state representation via deep neural networks allows Deep Reinforcement Learning (DRL) agents to achieve superior performance. Also it enables development of performant radiology models [4, 5, 6]. However, the models learned do not provide safety guarantees and are hard to analyze, which hinders their use in safety-critical applications.

An effective recent approach, called Viper, follows the DAgger imitation learning procedure [7] to create a decision tree model mimicking a DRL agent [8]. The key advantage of such decision tree models is that they are amenable to verification. Moreover, they are shown to perform well on environments such as Pong. However, decision trees are limited to axis perpendicular decision boundaries, which can adversely impact the performance. In this paper, we alleviate this issue by proposing a model with less restrictions on the geometry of decision boundaries.

We present MoËT (Mixture of Expert Trees), a technique based on Mixture of Experts (MoE) [9, 10, 11]. MoËT consists of decision tree (DT) experts and a gating function that determines the weights with which experts are used. Standard MoE models can typically use any expert as long as it is a differentiable function of model parameters. In this paper we tackle the problem of using non-differentiable decision trees in MoE context, as a means of obtaining verifiable DRL agents. Similar to MoE training by Expectation-Maximization (EM) algorithm, we first observe that MoËT can be trained by interchangeably optimizing the weighted log likelihood for experts (independently from one another) and optimizing the gating function with respect to the obtained experts. Based on that, we propose a procedure for DT learning in the specific context of MOE. To the best of our knowledge we are first to combine standard non-differentiable DT experts with MoE approach.

For a gating function, we use a simple generalized linear model with softmax function, which provides a distribution over experts. While decision boundaries of DTs are axis-perpendicular, the softmax gating function supports boundaries with hyperplanes of arbitrary orientations, thus improving

expressiveness. We also consider a variant of MoËT model that uses hard thresholding (MoËT$_{\text{h}}$) which selects just one most likely expert tree. Since MoE training algorithm tends to assign a region of space to a single expert ($P(e|r) \approx 1$) anyway, this variant does not suffer in performance, as we empirically demonstrate. Benefits of MoËT$_{\text{h}}$ compared to the soft version of MoËT are that it (a) allows for decomposing a decision into a set of logical rules, thus providing means for interpreting the model decisions, and (b) allows translation to satisfiability modulo theories (SMT) [1] formulas [12], thus providing rich opportunities for formal verification using off the shelf SMT solvers [2], as we demonstrate in the paper.

To employ MoËT in DRL setting we use the DAgger imitation learning procedure to mimic DRL agents. We evaluate our technique on six different environments: CartPole, Pong, Acrobot, Mountaincar, Lunarlander and Pendulum. We show that MoËT achieves better rewards and lower misprediction rates than Viper. Finally, we demonstrate how a MoËT policy for CartPole can be translated into an SMT formula to verify its properties using the Z3 theorem prover [13]. In addition we showed that MoËT can also be used in real-world supervised machine learning problems. We demonstrated that compared to the other verifiable machine learning models (logistic regression, decision trees and support vector classifiers with linear kernels) MoËT achieved much better results. By improving reliability of AI systems and to a degree improving their interpretability, our work aims at positive societal impact.

In summary, this paper makes the following key contributions:

1. We propose MoËT, a technique based on MoE with decision tree experts, and present a learning algorithm to train MoËT models.

2. We create MoËT$_{\text{h}}$, MoËT version with hard thresholding and softmax gating function which can be translated to an SMT formula amenable for verification and is not hard to interpret in case of small models.

3. We apply MoËT models in the RL setting, evaluate it on different environments and show that they lead to more performant models com-

[1] Very roughly, SMT is the problem of determining whether a mathematical formula is satisfiable, and it generalizes the Boolean satisfiability problem (SAT) to more complex formulas.

[2] SMT solvers are tools designed to solve SMT problems.

pared to Viper decision trees.

4. We apply MoËT models in real-world supervised problems and show that MoËT achieved better results compared to the others verifiable machine learning models.

The remainder of the paper is structured as follows. In section 2 the related work is reviewed. Motivating example to showcase some of the key difference between Viper and MoËT is presented in section 3, whereas background methodology is presented in 4. Explantion of MoËT model is given in section 5. Experimental setup and results obtained on different RL environments and supervised datasets are presented in section 6. The conclusions are drawn in section 7.

## 2. Related Work

**Verifiable Machine Learning:** RL algorithms are notoriously hard to debug and verify [14, 15]. A number of techniques has been proposed for enabling verification in RL setting [16, 17, 18, 19]. One existing approach synthesizes a program that approximates an RL policy [16]. The program acts as a shield, and their technique coordinates between using the shield program and original policy, which in combination provide safety guarantees. Instead of using a programmatic policy as a shield, another approach [18] creates a programmatic policy that can replace neural network policy altogether. Niu et al. [20] provide a general framework that leverages the success of verifiable and safe model-free RL in learning high performance controllers. Another system for verification of deep RL agents is presented in [17]. A hybrid RL agent framework that produces high-level autonomous verifiable behavior for unmanned vehicles is introduced in [21]. An abstraction approach, based on interval Markov decision processes, that yields probabilistic guarantees on accuracy of policy's execution, and presents techniques to build and solve different kind of control problems using abstract interpretation, mixed-integer linear programming, entropy-based refinement, and probabilistic model checking is presented in [22].

Compared to the other approaches, in this paper we propose a pure machine learning technique that is verifiable and applicable even outside of the RL setting. There has also been recent work on verification of random forests and tree ensembles [23, 24]. Such approaches might be useful in our future

work to extend verification from MoËT$_{\text{h}}$ to general MoËT models (which we dscribe later).

**Explainable Machine Learning:** There has been a lot of recent interest in explaining decisions of black-box models [25, 26, 27]. Nowadays, a large set of explainable RL litherature is emerging, intended to provide ethical, responsible and trustable algortihms for explaining model outputs of DRL agents [28, 29, 30]. Shi et al. [31] proposed XPM – an explainable RL framework for protfolio management optimization that is based on application of class activation mappings for output explanation. Similarly, Ayala et al. [32] proposed the introspection-based method for transforming Q-values into probabilities of success, used as the base to explain the agent's decision-making process. Besides of the explainable RL algorithms, the two most well known algorithms that are commonly used for deep learning models interpretation are LIME [33] and LORE [34]. LIME and LORE explain behaviour of a black-box model locally, around an input of interest, by sampling the black-box model around the neighborhood of the input, and training a local DT (or a linear model) over the sampled points.

Another view at MoËT is that it explains behavior of a deep RL agent. MoËT combines local trees into a global policy by combining local decision trees via a gating function. Inspection of the trees and the gating might shed light on the agent's decision making. However, we do not focus on this aspect in this paper.

**Tree-Structured Models:** Tree-Structured models are very attractive type of machine learning algorithms due to low complexity and interpretability [35, 36]. Irsoy et al. [37] propose a decision tree model with soft decisions at internal nodes where children are chosen with probabilities given by a sigmoid gating function. However, this reduces the tree's interpretability. Binary tree-structured hierarchical routing mixture of experts (HRME) model, which has classifiers as non-leaf node experts and simple regression models as leaf node experts, was proposed in [38]. Hester and Stone [39] use random forests in RL setting to build a model of environment from which policy is inferred.

The form of our model can be related to these models, but it is designed with verifiability in mind and we also propose a novel training procedure suited to that specific model.

**Knowledge Distillation and Model Compression:** We rely on ideas already explored in fields of model compression [40] and knowledge distillation [41, 42, 43]. The idea is to use a complex well performing model to

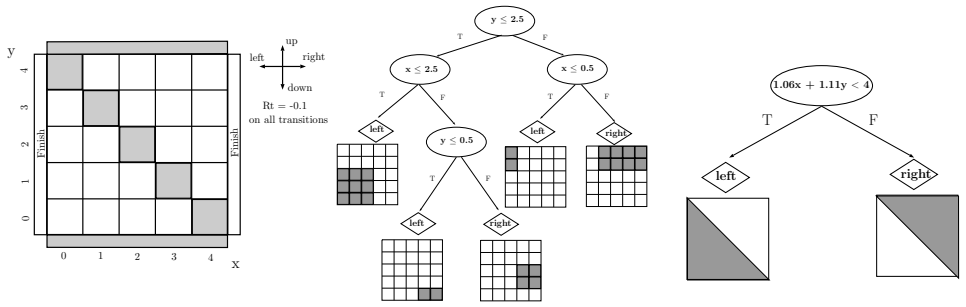


Figure 1: An illustratory Gridworld environment (left), a Viper policy learned for the environment (middle), and a MoËT policy learned for the environment (right).

facilitate training of a simpler model which might have some other desirable properties (e.g., verifiability and interpretability). Such practices have been applied to approximate decision tree ensemble by a single tree [44]. In contrast, we approximate a neural network. Similarly, a neural network can be used to train another neural network [45], but neural networks are hard to interpret and even harder to formally verify. Such practices have also been applied in the field of reinforcement learning in knowledge and policy distillation [46, 47, 48, 49, 50], which are similar in spirit to our work, and imitation learning [8, 7, 51, 52], which provide a foundation for our work.

## 3. Motivating Example: Gridworld

We now present a simple motivating example to showcase some of the key differences between Viper and MoËT approaches. Consider the $N \times N$ Gridworld problem shown in Figure 1 (for $N = 5$). The agent is placed at a random position in a grid (except the walls denoted by filled rectangles) and should find its way out. To move through the grid the agent can choose to go up, left, right or down at each time step. If it hits the wall (gray cell) it stays in the same position (state). State is represented using two integer values ($x, y$ coordinates) which range from $(0, 0)$—bottom left to $(N - 1, N - 1)$—top right. The grid can be escaped through either left doors (left of the first column), or right doors (right of the last column). A negative reward of $-0.1$ is received for each agent action (negative reward encourages the agent to find the exit as fast as possible). An episode finishes as soon as an exit is reached or if 100 steps are made whichever comes first.

Table 1: Size comparison of MoËT and Viper DT policies on the Gridworld problem (Figure 1), for different sizes of the square board ($N \times N$). The left side of the table presents the depths of obtained models (that perfectly mimic optimal policy) for MoËT and for Viper (DTs), while the right side presents the number of nodes in these models. Both the depth and the number of nodes show that by increasing size of the grid ($N$) size of MoËT models stays constant, while Viper (DT) models grow in size.

| | **Depth** | | **Nodes** | |
|---|---|---|---|---|
| **N** | **MoËT** | **Viper DT** | **MoËT** | **Viper DT** |
| 5 | 1 | 3 | 3 | 9 |
| 6 | 1 | 4 | 3 | 11 |
| 7 | 1 | 4 | 3 | 13 |
| 8 | 1 | 4 | 3 | 15 |
| 9 | 1 | 4 | 3 | 17 |
| 10 | 1 | 5 | 3 | 21 |

The optimal policy ($\pi_*$) for this problem consists of taking the left (right resp.) action for each state below (above resp.) the diagonal. We used $\pi_*$ as a teacher and imitation learning approach of Viper to train an interpretable DT policy that mimics $\pi_*$. The resulting DT policy is shown in Figure 1. The DT partitions the state space (grid) using lines perpendicular to x and y axes, until it separates all states above diagonal from those below. This results in a DT of depth 3 with 9 nodes. On the other hand, the policy learned by MoËT is shown in Figure 1. The MoËT model with 2 experts learns to partition the space using the line defined by a linear function $1.06x + 1.11y = 4$ (roughly the diagonal of the grid). Points on the different sides of the line correspond to two different experts which are themselves DTs of depth 0 always choosing to go left (below) or right (above).

We notice that DT policy needs much larger depth to represent $\pi_*$ while MoËT can represent it as only one decision step. Furthermore, with increasing $N$ (size of the grid), complexity of DT grows, while MoËT complexity stays the same; we empirically confirm this as follows. For Gridworld sizes $N = 5, 6, 7, 8, 9, 10$, the depths of obtained DTs are $3, 4, 4, 4, 4, 5$ and the numbers of their nodes are $9, 11, 13, 15, 17, 21$ respectively. In contrast, MoËT models of the same complexity and structure as the one shown in Figure 1 are learned for all values of $N$. We present these results in Table 1 for better readability (all policies learned are equivalent to $\pi_*$).

## 4. Background

In this section we provide description of two relevant methods we build upon: (1) Viper, an approach for interpretable imitation learning, and (2) MoE learning framework.

**Viper**. Viper algorithm (included in appendix) is an instance of DAGGER imitation learning approach, adapted to prioritize critical states based on Q-values. Inputs to the Viper training algorithm are (1) environment $e$ which is an finite horizon ($T$-step) Markov Decision Process (MDP) $(S, A, P, R)$ with states $S$, actions $A$, transition probabilities $P : S \times A \times S \to [0, 1]$, and rewards $R : S \to \mathbb{R}$; (2) teacher policy $\pi_t : S \to A$; (3) its Q-function $Q^{\pi_t} : S \times A \to \mathbb{R}$ and (4) number of training iterations $N$. Distribution of states after $T$ steps in environment $e$ using a policy $\pi$ is $d^{(\pi)}(e)$ (assuming randomly chosen initial state). Viper uses the teacher as an oracle to label the data (states with actions). It initially uses teacher policy to sample trajectories (states) to train a student (DT) policy. It then uses the student policy to generate more trajectories. Viper samples training points from the collected dataset $D$ giving priority to states $s$ having higher importance $I(s)$, where $I(s) = \max_{a \in A} Q^{\pi_t}(s, a) - \min_{a \in A} Q^{\pi_t}(s, a)$. This sampling of states leads to faster learning and shallower DTs. The process of sampling trajectories and training students is repeated for number of iterations $N$, and the best student policy is chosen using reward as the criterion.

**Mixture of Experts**. MoE is an ensemble model [9, 10, 11] that consists of expert networks and a gating function. Gating function divides the input (feature) space into regions for which different experts are specialized and responsible. MoE is flexible with respect to the choice of expert models as long as they are differentiable functions of model parameters (which is not the case for DTs).

In MoE framework, probability of outputting $\mathbf{y} \in \mathbb{R}^m$ given an input $\mathbf{x} \in \mathbb{R}^n$ is given by:

$$P(\mathbf{y}|\mathbf{x}, \theta) = \sum_{i=1}^{E} P(i|\mathbf{x}, \theta_g) P(\mathbf{y}|\mathbf{x}, \theta_i) = \sum_{i=1}^{E} g_i(\mathbf{x}, \theta_g) P(\mathbf{y}|\mathbf{x}, \theta_i) \qquad (1)$$

where $E$ is the number of experts, $g_i(\mathbf{x}, \theta_g)$ is the probability of choosing the expert $i$ (given input $\mathbf{x}$), $P(\mathbf{y}|\mathbf{x}, \theta_i)$ is the probability of expert $i$ producing output $\mathbf{y}$ (given input $\mathbf{x}$). Learnable parameters are $\theta = (\theta_g, \theta_e)$, where $\theta_g$ are parameters of the gating function and $\theta_e = (\theta_1, \theta_2, ..., \theta_E)$ are parameters

of the experts. Gating function can be modeled using a softmax function over a set of linear models. Let $\theta_g$ consist of parameter vectors $(\theta_{g1}, \ldots, \theta_{gE})$, then the gating function can be defined as $g_i(\mathbf{x}, \theta_g) = {\exp(\theta_{gi}^T \mathbf{x})}/{\sum_{j=1}^{E} \exp(\theta_{gj}^T \mathbf{x})}$ .

In the case of classification, an expert $i$ outputs a vector $\mathbf{y}_i$ of length $C$, where $C$ is the number of classes. Expert $i$ associates a probability to each output class $c$ (given by $\mathbf{y}_{ic}$) using the gating function. Final probability of a class $c$ is a gate weighted sum of $\mathbf{y}_{ic}$ for all experts $i \in 1, 2, ..., E$. This creates a probability vector $\mathbf{y} = (y_1, y_2, ..., y_C)$, and the output of MoE is $\arg\max_i \mathbf{y}_i$.

MoE is commonly trained using an EM algorithm, where instead of direct optimization of the likelihood one performs optimization of an auxiliary function $\hat{L}$ defined in a following way. Let $z$ denote the expert chosen for instance $\mathbf{x}$. Then joint likelihood of $\mathbf{x}$ and $z$ can be considered. Since $z$ is not observed in the data, log likelihood of samples $(\mathbf{x}, z, \mathbf{y})$ cannot be computed, but instead expected log likelihood can be considered, where expectation is taken over $z$. Since the expectation has to rely on some distribution of $z$, in the iterative process, the distribution with respect to the current estimate of parameters $\theta$ is used. More precisely function $\hat{L}$ is defined by [10]:

$$\hat{L}(\theta, \theta^{(k)}) = \mathbb{E}_z[\log P(\mathbf{x}, z, \mathbf{y})|\mathbf{x}, \mathbf{y}, \theta^{(k)}] = \int P(z|\mathbf{x}, \mathbf{y}, \theta^{(k)}) \log P(\mathbf{x}, z, \mathbf{y}) dz \tag{2}$$

where $\theta^{(k)}$ is the estimate of parameters $\theta$ in iteration $k$. Then, for a specific sample $D = \{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, N\}$, the following formula can be derived [10]:

$$\hat{L}(\theta, \theta^{(k)}) = \sum_{i=1}^{N} \sum_{j=1}^{E} h_{ij}^{(k)} \log g_j(\mathbf{x}_i, \theta_g) + \sum_{i=1}^{N} \sum_{j=1}^{E} h_{ij}^{(k)} \log P(\mathbf{y}_i|\mathbf{x}_i, \theta_j) \tag{3}$$

where it holds

$$h_{ij}^{(k)} = \frac{g_j(\mathbf{x}_i, \theta_g^{(k)}) P(\mathbf{y}_i|\mathbf{x}_i, \theta_j^{(k)})}{\sum_{l=1}^{E} g_l(\mathbf{x}_i, \theta_g^{(k)}) P(\mathbf{y}_i|\mathbf{x}_i, \theta_l^{(k)})} \tag{4}$$

## 5. Mixture of Expert Trees

In this section we explain the adaptation of original MoE model to mixture of decision trees, and present both training and inference algorithms.

Considering that coefficients $h_{ij}^{(k)}$ (Eq. 4) are fixed with respect to $\theta$ and that in Eq. 3 the gating part (first double sum) and each expert part depend on disjoint subsets of parameters $\theta$, training can be carried out by interchangeably optimizing the weighted log likelihood for experts (independently from one another) and optimizing the gating function with respect to the obtained experts. The training procedure for MoËT, described by Algorithm 1, is based on this observation. First, the parameters of the gating function are randomly initialized (line 2). Then the experts are trained one by one. Each expert $j$ is trained on a dataset $D_w$ of instances weighted by coefficients $h_{ij}^{(k)}$ (line 5), by applying specific DT learning algorithm (line 6) that we adapted for MoE context (described below). After the experts are trained, an optimization step is performed (line 7) in order to increase the gating part of Eq. 3. At the end, the parameters are returned (line 8).

Our tree learning procedure is as follows. Our technique modifies original MoE algorithm in that it uses DTs as experts. The fundamental difference with respect to traditional model comes from the fact that DTs do not rely on explicit and differentiable loss function which can be trained by gradient descent or Newton's methods. Instead, due to their discrete structure, they rely on a specific greedy training procedure. Therefore, the training of DTs has to be modified in order to take into account the attribution of instances to the experts given by coefficients $h_{ij}^{(k)}$, sometimes called *responsibility* of expert $j$ for instance $i$. If these responsibilities were hard, meaning that each instance is assigned to strictly one expert, they would result in partitioning the feature space into disjoint regions belonging to different experts. On the other hand, soft responsibilities are fractionally distributing each instance to different experts. The higher the responsibility of an expert $j$ for an instance $i$, the higher the influence of that instance on that expert's training. In order to formulate this principle, we consider which way the instance influences construction of a tree. First, it affects the impurity measure computed when splitting the nodes and second, it influences probability estimates in the leaves of the tree. We address these two issues next.

A commonly used impurity measure to determine splits in the tree is the Gini index. Let $U$ be a set of indices of instances assigned to the node for which the split is being computed and $D_U$ set of corresponding instances. Let categorical outcomes of $y$ be $1, \ldots, C$, and for $l = 1, \ldots, C$ let denote $p_l$ as a fraction of instances in $D_U$ for which it holds $y = l$. More formally $p_l = \frac{\sum_{i \in U} I[y_i = l]}{|U|}$, where $I$ denotes indicator function of its argument expression

**Algorithm 1** MoËT training.

1: **procedure** MoËT (Data $\{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, N\}$, Epochs $N_E$, Number of Experts $E$)
2: $\quad\theta_{\mathbf{g}} \leftarrow initialize()$
3: $\quad$**for** $k \leftarrow 1$ to $N_E$ **do**
4: $\quad\quad$**for** $j \leftarrow 1$ to $E$ **do**
5: $\quad\quad\quad D_w \leftarrow \left\{ \left( \mathbf{x}_i, \mathbf{y}_i, \frac{g_j(\mathbf{x}_i,\theta_g)P(\mathbf{y}_i|\mathbf{x}_i,\theta_j)}{\sum_{e=1}^{E} g_e(\mathbf{x}_i,\theta_g)P(\mathbf{y}_i|\mathbf{x}_i,\theta_e)} \right) \mid i = 1, \ldots, N \right\}$
6: $\quad\quad\quad \theta_j \leftarrow fit\_tree(D_w)$
7: $\quad\quad \theta_g \leftarrow \theta_g + \lambda \nabla_{\theta'} \sum_{i=1}^{N} \sum_{j=1}^{E} \left[ \frac{g_j(\mathbf{x}_i,\theta_g)P(\mathbf{y}_i|\mathbf{x}_i,\theta_j)}{\sum_{e=1}^{E} g_e(\mathbf{x}_i,\theta_g)P(\mathbf{y}_i|\mathbf{x}_i,\theta_e)} \log g_j(\mathbf{x}_i, \theta') \right]$
8: $\quad$**return** $\theta_g, (\theta_1, \ldots, \theta_E)$

and equals 1 if the expression is true. Then the Gini index $G$ of the set $D_U$ is defined by: $G(p_1, \ldots, p_C) = 1 - \sum_{l=1}^{C} p_l^2$. Considering that the assignment of instances to experts are fractional as defined by responsibility coefficients $h_{ij}^{(k)}$ (which are provided to tree fitting function as weights of instances computed in line 5 of the algorithm), this definition has to be modified in that the instances assigned to the node should not be counted, but instead, their weights should be summed. Hence, we propose the following definition:

$$\hat{p}_l = \frac{\sum_{i \in U} I[y_i = l] h_{ij}^{(k)}}{\sum_{i \in U} h_{ij}^{(k)}} \tag{5}$$

and compute the Gini index for the set $D_U$ as $G(\hat{p}_1, \ldots, \hat{p}_C)$. Similar modification can be performed for other impurity measures (such as entropy) relying on distribution of outcomes of a categorical variable. Note that while the instance assignments to experts are soft, instance assignments to nodes within an expert are hard, meaning sets of instances assigned to different nodes are disjoint. Probability estimate for $\mathbf{y}$ in the leaf node is usually performed by computing fractions of instances belonging to each class. Instead of such estimates, again, we use estimates $\hat{p}_l$ defined by Eq. 5. Hence, the estimates of probabilities $P(\mathbf{y}|\mathbf{x}, \theta_j^{(k)})$ needed by MoE are defined. In Algorithm 1, function $fit\_tree$ performs decision tree training using the above modifications.

We consider two ways to perform inference with respect to the obtained model. First one which we call MoËT, is performed by maximizing $P(\mathbf{y}|\mathbf{x}, \theta)$ with respect to $\mathbf{y}$ where this probability is defined by Eq. 1. The second way,

which we call MoËT$_{\text{h}}$, performs inference as $\arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \theta_{\arg\max_j g_j(x,\theta_g)})$, meaning that we only rely on the most probable expert.

**Adaptation of MoËT to imitation learning**. We integrate MoËT model into imitation learning approach of Viper by substituting training of DT with the MoËT training procedure.

**Verifiability by translating MoËT to SMT**. We define a translation of MoËT$_{\text{h}}$ models to SMT formulas, which opens a range of possibilities for validating and interpreting the model using automated reasoning tools. SMT formulas provide a rich means of logical reasoning, where a user can query the solver with questions such as: "What inputs do the two models differ on?", or "What is the closest input to the given input using which model makes a different prediction?", or "Are the two models equivalent?", or "Are the two models equivalent in respect to the output class C?". Answers to such questions can help better understand and compare models in a rigorous way. Also note that the symbolic reasoning of the gating function and decision trees allows construction of SMT formulas that are readily handled by off-the-shelf tools, whereas direct SMT encoding of neural networks do not scale for any reasonably sized network because of the need for non-linear arithmetic reasoning.

We show the translation of MoËT policy to SMT constraints for verifying policy properties. We present an example translation of MoËT policy on CartPole environment with the same property specification that was proposed for verifying Viper policies [8]. The goal in CartPole is to keep the pole upright, which can be encoded as a formula:

$$\psi \equiv s_0 \in S_0 \wedge \bigwedge_{t=1}^{\infty} |\phi(f_t(s_{t-1}, \pi(s_{t-1}))| \leq y_0$$

where $s_i$ represents state after $i$ steps, $\phi$ is the deviation of pole from the upright position. In order to encode this formula it is necessary to encode the transition function $f_t(s, a)$ which models environment dynamics: given a state and action it returns the next state of the environment. Also, it is necessary to encode the policy function $\pi(s)$ that for a given state returns action to perform. There are two issues with verifying $\psi$: (1) infinite time horizon; and (2) the nonlinear transition function $f_t$. To solve this problem, Bastani et al. [8] use a finite time horizon $T_{max} = 10$ and linear approximation of the dynamics. We make the same assumptions.

To encode $\pi(s)$ we need to translate both the gating function and DT

experts to logical formulas. Since the gating function in MoËT$_{\text{h}}$ uses exponential function, it is difficult to encode the function directly in Z3 as SMT solvers do not have efficient decision procedures to solve non-linear arithmetic. The direct encoding of exponentiation therefore leads to prohibitively complex Z3 formulas. We exploit the following simplification of the gating function that is sound when hard prediction is used:

$$e = \arg\max_i \left( \frac{\exp(\theta_{gi}^T \mathbf{x})}{\sum_{j=1}^{E} \exp(\theta_{gj}^T \mathbf{x})} \right) = \arg\max_i (\exp(\theta_{gi}^T \mathbf{x})) = \arg\max_i (\theta_{gi}^T \mathbf{x}) \quad (6)$$

First simplification is possible since the denominators of the gating functions are same for all experts, and second is due to the monotonicity of the exponential function. We use the same DT encoding as in Viper. To verify that $\psi$ holds we need to show that $\neg\psi$ is unsatisfiable. In the experimental evaluation we run the verification with our MoËT$_{\text{h}}$ policies and show that $\neg\psi$ is indeed unsatisfiable.

**Expressiveness**. DTs make their decisions by partitioning the feature space into regions which have borders perpendicular to coordinate axes. To approximate borders that are not perpendicular to coordinate axes very deep trees are often necessary. MoËT$_{\text{h}}$ mitigates this shortcoming by exploiting hard softmax partitioning of the feature space using borders which are still hyperplanes, but need not be perpendicular to coordinate axes (see Section 3), which improves the expressiveness.

**Interpretability**. While we do not focus on interpretability in this work, it is useful to note that MoËT$_{\text{h}}$ models do exhibit some interpretability properties. A MoËT$_{\text{h}}$ model is a combination of a linear model and several decision tree models. Only a single DT is used for each prediction (instead of weighted average), which facilitates interpretability. If the models are small (e.g, depth $\leq 10$) and include small number of features, a person can easily simulate and understand the model. This observations resonate with several points about interpretability made in [53]

**Limitations**. Our work tries to strike a balance between expressiveness, which allows for more performant models, and verifiability, which allows for more reliable models. Therefore, while being more expressive than decision trees, MoËT still has limited expressiveness compared to deep learning models, which is a price paid for easier verifiability.

## 6. Evaluation

We first discuss DRL agents we use as a starting point in the imitation learning. Second, we explore the performance capabilities of Viper by finding decision tree depths at which the performance saturates—cannot be improved by increasing the depth further. Then, after ensuring that we explored the useful space of configurations for Viper, we pick the best performing Viper models and compare them with the best performing MoËT models to quantitatively compare the two. Finally, we re-evaluate performance of the models to evaluate how well they generalize. Also, we verify $\text{MoË}\text{T}_\text{h}$ policies on CartPole environment and visually compare the expressiveness of different policies. Eventually, we presented that MoËT can be also successfully applied in real-world supervised learning problems.

**DRL agents**. We use following OpenAI Gym environments in our evaluation: CartPole, Acrobot, Mountaincar, Lunarlander, Pong and Pendulum (description of the environments is included in the appendix). For DRL agents, we use a policy gradient model in CartPole, a deep Q-network (DQN) [54] in Pong, and dueling DQN [55] in the other environments (training hyperparameters provided in the appendix). We train MoËT and Viper policies by mimicking the agents. The rewards (total return during an episode) obtained by the DRL agents on CartPole, Acrobot, Mountaincar, Lunarlander, Pong and Pendulum are 200.00, −68.60, −105.27, 190.90, 21.00 and −158.13, respectively. Rewards are averaged across 100 (250 in CartPole) runs (episodes).

**Performance saturation of Viper**. We first examine performance capabilities of Viper, i.e., answer the question of when the performance saturates, by examining performance of decision trees of gradually increased maximum depth (Figure 2). For each depth we train multiple Viper models and show performance trends in terms of reward and fidelity. By reward we mean cumulative reward achieved during an episode, while fidelity represents percent of times a student performs the same action as its teacher (DRL agent). Achieving high reward indicates that a student is performing well, while high fidelity indicates that the student policy is close to the teacher's. We ensure to train at least 5 different Viper models for each depth.[3] Us-

[3] We train at least 5 Viper models for each subject and maximum depth value. Due to the computational limitations actual number of Viper models trained varies across environments: CartPole $\in [35, 70]$, Acrobot $\in [10, 70]$, Mountaincar $\in [10, 70]$, Lunarlander

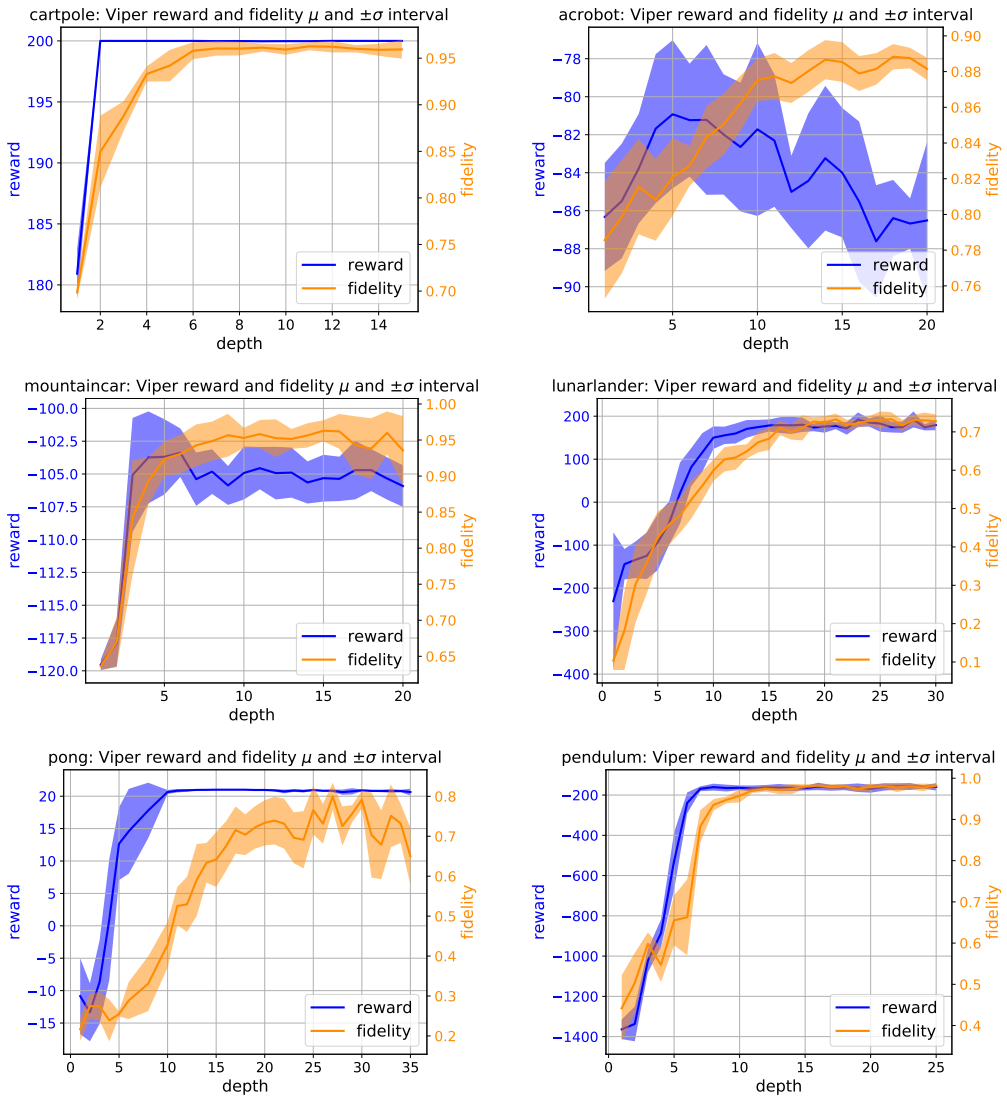


Figure 2: **Performance saturation of Viper**. Multiple models are trained for a single maximum depth of Viper decision trees, while maximum depth is incrementally increased, showing the mean value and standard deviation of reward and fidelity in respect to the depth. These results inform when Viper performance saturates, i.e., reaches a point upon which increasing maximum depth is not helpful anymore, we call that point performance saturation depth.

$\in [10, 70]$, Pong $\in [5, 24]$ and Pendulum $\in \{10\}$.

ing the performance trend plots we infer when Viper performance saturates, i.e., reaches a depth after which further increasing maximum depth does not help. Performance saturation depths for CartPole, Acrobot, Mountaincar, Lunarlander, Pong and Pendulum are 8, 15, 12, 20, 30 and 20, respectively. Identifying the performance saturation points for Viper is helpful in identifying the overall best performing Viper model, thus giving confidence during comparison with MoËT models that we explored the useful space of Viper configurations.

**Best performing Viper, MoËT and MoËT$_{\mathbf{h}}$ models**. We next compare Viper, MoËT and MoËT$_{\text{h}}$ models by visualizing their Pareto fronts with respect to the reward and fidelity (Figure 3). Pareto front of a set of models consists of all models from that set which are not dominated by any other model from the set in terms of reward or fidelity. In other words, every model dominated by another model in terms of both metrics is not considered. From the set of all Viper models trained for different maximum depths (from depth 1 to the saturation depth) we select models on the Pareto front. Similar is done for MoËT and MoËT$_{\text{h}}$ which we trained for different number of experts and expert depths (information about configurations used is provided in the appendix). A global Pareto front (best models across all architectures) is shown with points connected by a black solid line.

By inspecting the results we notice that in the case of CartPole, all 3 models achieve maximum reward (200), however fidelity is significantly higher in the case of MoËT and MoËT$_{\text{h}}$ (over 99% compared to 97%). Also, it is interesting to note that both MoËT and MoËT$_{\text{h}}$ models on the Pareto front consist of 2 experts of depth 0, while the Viper model on the Pareto front is a decision tree of depth 6. In the case of Acrobot, we notice that MoËT models dominate MoËT$_{\text{h}}$ and Viper models, and that MoËT$_{\text{h}}$ models dominate Viper models. Thus, both MoËT and MoËT$_{\text{h}}$ models achieve higher reward and fidelity over Viper models. In the case of Mountaincar, the global Pareto front contains some Viper models, but mostly MoËT and MoËT$_{\text{h}}$ dominate. Furthermore, models exhibiting the highest reward as well as fidelity are MoËT and MoËT$_{\text{h}}$ models. In the case of Lunarlander, both MoËT and MoËT$_{\text{h}}$ dominate Viper models. A MoËT$_{\text{h}}$ model achieves the maximum reward of over 260 while a Viper model achieves the maximum reward of around 215. Furthermore, both MoËT and MoËT$_{\text{h}}$ models achieve better fidelity compared to Viper. In the case of Pong, all 3 models achieve maximum reward (21), however fidelity is higher for MoËT and MoËT$_{\text{h}}$. In the case of Pendulum, MoËT and MoËT$_{\text{h}}$ models achieve

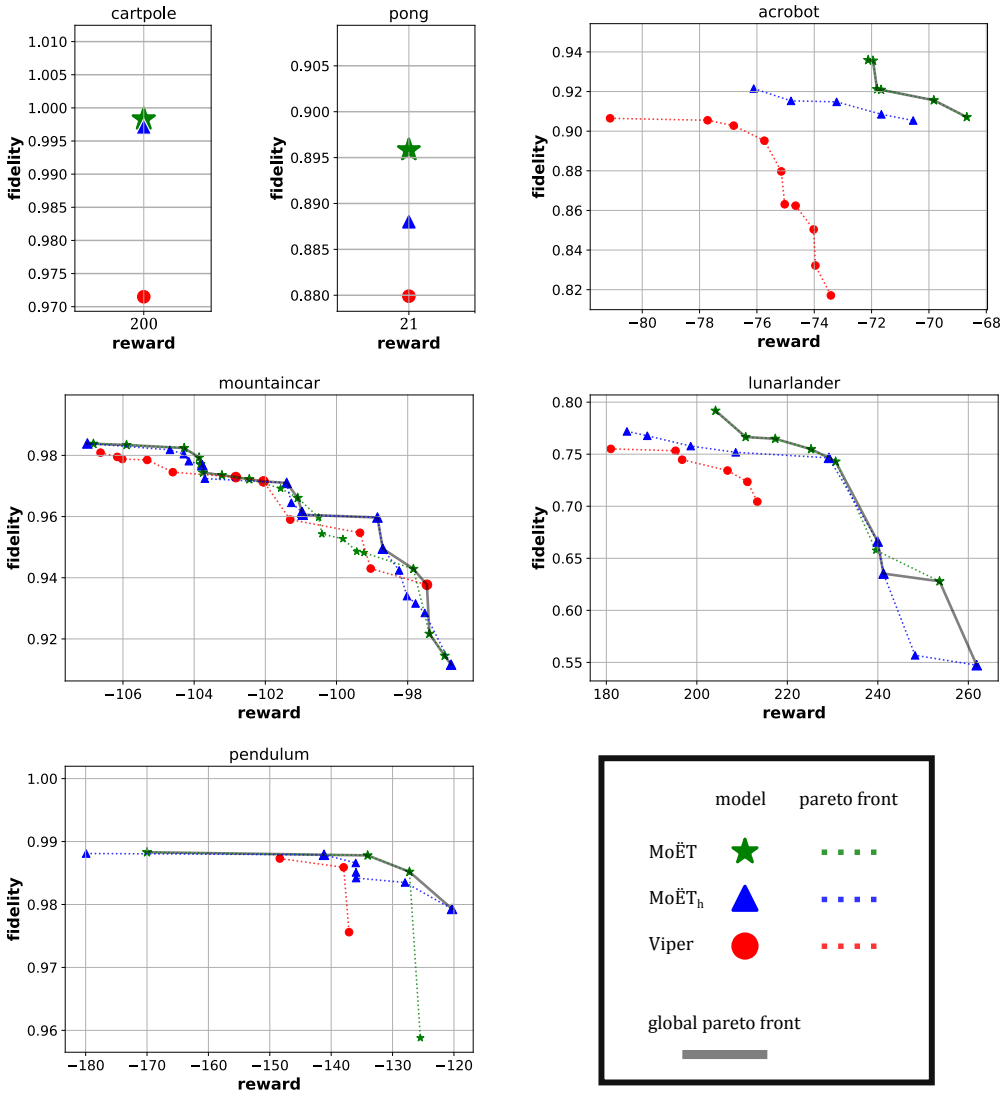


Figure 3: **Best performing Viper, MoËT and MoËT$_h$ models**. Pareto fronts (in respect to the reward and fidelity) are identified separately for Viper, MoËT and MoËT$_h$ models. Global Pareto fronts are shown with points connected by a gray solid line.

better maximum reward, while maximum fidelity is about equal for all the models. Note that for a given fidelity score, MoËT and MoËT$_h$ are advantageous to Viper. Scores of the points on the global Pareto front are presented in a tabular form in Appendix E.

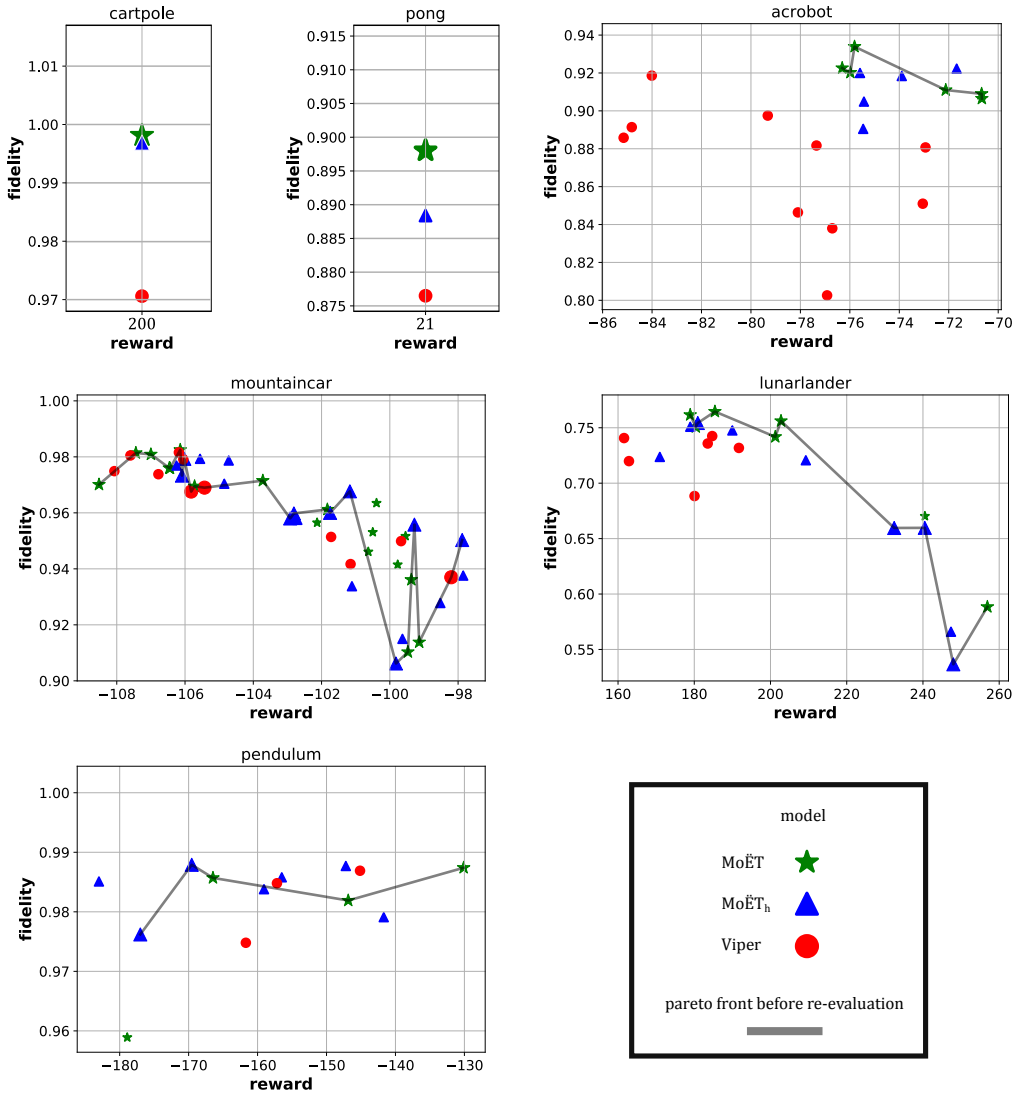


Figure 4: **Performance generalization of models**. Models on the Pareto fronts (Figure 3) are re-evaluated. Black solid line connects models that were on the global Pareto front before re-evaluation.

**Performance generalization of models**. In the supervised learning setting, after the best models are selected based on their performance on a validation set, they are re-evaluated on a test set to get a better estimate of their performance on the new data. In RL setting there is no direct analogy

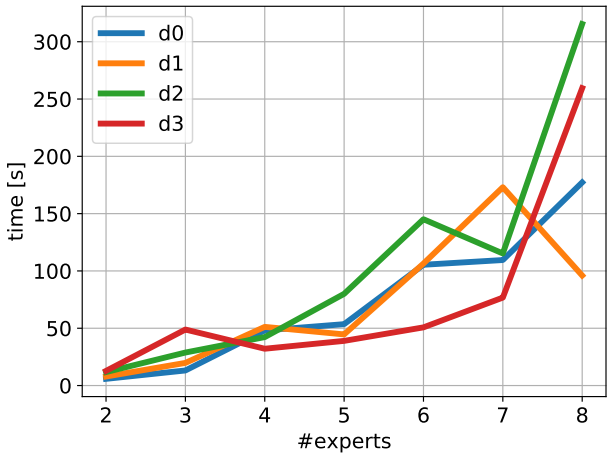


Figure 5: Verification times.

to validation and test datasets, but the models can be re-evaluated after the selection is performed. After we identify the best models on the Pareto fronts (Figure 3), we re-evaluate their performance by running them again through the RL environment. Figure 4 shows the achieved performance of these models after re-evaluation. In the case of CartPole and Pong performance before and after re-evaluation are very similar. In the case of Acrobot, Mountain-car and Lunarlander, models that were on the global Pareto front are mostly still on the global Pareto front in the re-evaluation. Moreover, MoËT and MoËT$_{\text{h}}$ models dominate Viper models in most of the cases. Pendulum environment behaves more stochastically – evaluating policy (done across 100 episodes) can exhibit significantly different reward from evaluation to evaluation, making results more inconclusive. However, all models achieve great fidelity level, and reward that is close to the DRL agent one. Considering high performance, differences in performance between models are minor. Scores of the points that were on the global Pareto front are presented in a tabular form in Appendix E.

Following the previous analysis, we conclude that MoËT and MoËT$_{\text{h}}$ models provide better performance (in terms of reward and fidelity) compared to Viper in most of the cases, demonstrating that MoËT is a valuable technique to be considered when looking for a verifiable RL policy.

**Verification**. We perform verification of MoËT$_{\text{h}}$ policies obtained in

our experiments according to the procedure described in Section 5. All models considered in this experiment successfully pass the verification procedure. To better understand the scalability of our verification procedure, we report the verification times needed to verify policies for different number of experts and expert depths in Figure 5. The verification times generally increase with the number of experts. MoËT$_{\text{h}}$ policies with 2 experts take from 5.5s to 11.7s for verification, while the verification times for 8 experts can go up to as much as 336s. This corresponds to the complexity of the logical formula obtained with an increase in the number of experts. While the effect of expert depths on verification times is visible in a case of few experts, with the increase of experts it is less noticeable, thus indicating that the number of experts has more influence on the verification times than expert depths. We run the verification on Intel i7-7600, 2.80GHz, 16 GB LPDDR3. We show example SMT formula (of Viper and MoËT$_{\text{h}}$ policies) in Appendix D.

**Expressiveness**. We provide a simple qualitative comparison of best Viper and MoËT$_{\text{h}}$ policies, by contrasting them to DRL policy on a Cart-Pole environment. The figure 6 visualizes these policies and demonstrates that MoËT$_{\text{h}}$ policy much more closely resembles the DRL policy thanks to its ability to represent hyperplanes of arbitrary orientation, while DT policy obtained by Viper approximates DRL policy by axis perpendicular hyperplanes. The MoËT$_{\text{h}}$ policy presented is equivalent to the following program: `if` $2.18 * cp + 7.22 * cv + 20.64 * pa + 25.33 * pv > -1$ `then` go right `else` go left, where $cp$ and $cv$ are cart position and velocity, and $pv$ and $pa$ pole angle and its angular velocity.

**Supervised learning.** We evaluated the performance of MoËT and MoËT$_{\text{h}}$ in the supervised regime on three real-world datasets. Two datasets (German credit and Adult income) come from the UCI ML repository [56], whereas the Fetal health dataset is a publicly available dataset that can be found on Kaggle. We summarize the properties of the datasets that we use in Table 2.

In the *Adult income* dataset [57] the goal is to predict whether an income is greater than 50K dollars. In the *German credit* dataset, the goal is to classify bank account holders into two classes – good or bad. In the *Fetal health* dataset, the goal is to predict whether a fetus is healthy or not based on the features extracted from cardiotocogram examination.

We compared MoËT with other supervised learning models which would require similar effort and tools to be verified: decision tree, support vector classifier (SVC) with linear kernel, ridge logistic regression and lasso logistic

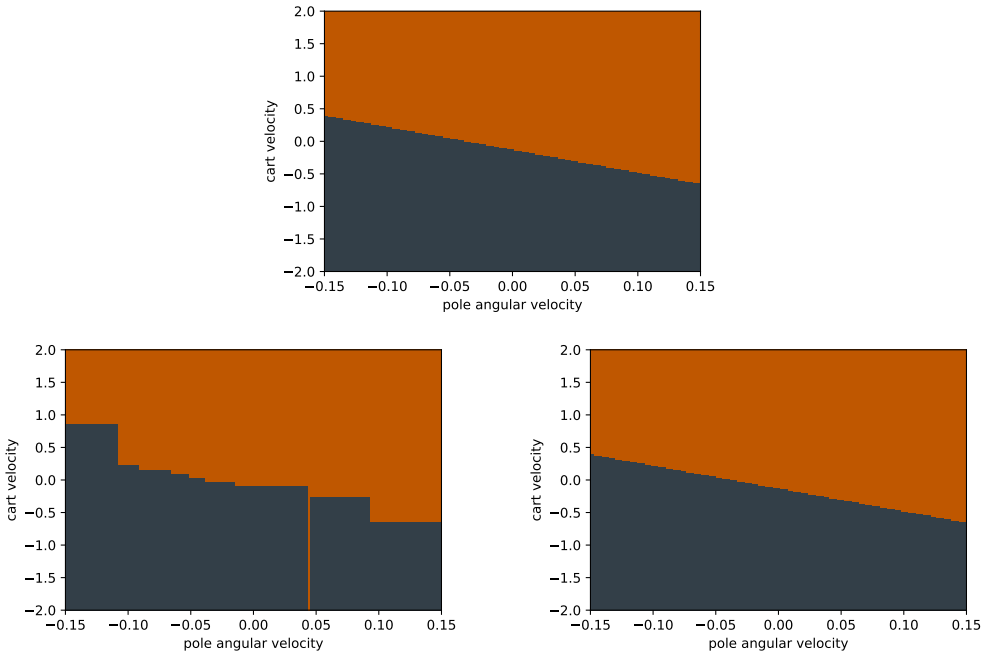


Figure 6: **Visualizing DRL (top), Viper (bottom left) and MoËT$_{\mathbf{h}}$ (bottom right) policies on CartPole**. X-axis represent pole angular velocity and y-axis cart velocity, which are the most discriminatory features (topmost nodes in the Viper decision tree policy). Other features, cart position and pole angle, are set to 0 (center position with pole upright). Gray color represents points where agent takes action *left*, and orange points when agent takes action *right*.

regression. The results are evaluated by F1 score and accuracy. The hyperparameters of compared models are tuned on validation set. The results evaluated on test set with 95% confidence intervals for *Fetal health*, *German credit*, and *Adult income* datasets are presented in Tables 3, 4, and 5, respectively. It can be observed that MoËT is the best performing model with exception of SVC being better on German credit data according to accuracy (but not F1 score). Therefore, it can be concluded that MoËT can also be successfully applied in the case of supervised learning problems.

## 7. Conclusion

We introduced MoËT, a technique based on MoE with decision trees as experts and formulated a learning algorithm to train MoËT models. To

Table 2: For each dataset used in the experimental evaluation we provide its name, the number of instances it contains (Size), numbers of instances per set after splitting the data into training, validation, and testing sets (Split) and total number of features (Features)

| **Dataset** | **Size** | **Split (train/test/val)** | **Features** |
|---|---|---|---|
| Adult income | 48,842 | 34,189 / 6,783 / 6,784 | 14 |
| German credit | 1,000 | 700 / 150 / 150 | 10 |
| Fetal health | 2126 | 1488 / 319 / 319 | 21 |

Table 3: Prediction performance of classifiers - *Fetal health* dataset

| **model/metrics** | **F1 score** | **Accuracy** |
|---|---|---|
| Decision tree | 0.852 ± 0.004 | 0.939 ± 0.004 |
| Lasso logistic regression | 0.797 ± 0.000 | 0.915 ± 0.000 |
| MoËT$_{\text{h}}$ | 0.880 ± 0.001 | 0.950 ± 0.001 |
| MoËT | **0.891 ± 0.001** | **0.955 ± 0.001** |
| Ridge logistic regression | 0.739 ± 0.000 | 0.903 ± 0.000 |
| SVC | 0.762 ± 0.000 | 0.906 ± 0.000 |

Table 4: Prediction performance of classifiers - *German credit* dataset

| **model/metrics** | **F1 score** | **Accuracy** |
|---|---|---|
| Decision tree | 0.759 ± 0.000 | 0.637 ± 0.000 |
| Lasso logistic regression | 0.797 ± 0.000 | 0.667 ± 0.000 |
| MoËT$_{\text{h}}$ | 0.759 ± 0.003 | 0.638 ± 0.004 |
| MoËT | **0.808 ± 0.003** | 0.687 ± 0.004 |
| Ridge logistic regression | 0.792 ± 0.000 | 0.660 ± 0.000 |
| SVC | 0.799 ± 0.000 | 0.**693 ± 0.000** |

Table 5: Prediction performance of classifiers - *Adult income* dataset

| model/metrics | F1 score | Accuracy |
|---|---|---|
| Decision tree | 0.661 ± 0.003 | 0.852 ± 0.001 |
| Lasso logistic regression | 0.536 ± 0.000 | 0.820 ± 0.000 |
| MoËT$_{\text{h}}$ | **0.676 ± 0.000** | 0.854 ± 0.000 |
| MoËT | **0.674 ± 0.004** | **0.860 ± 0.001** |
| Ridge logistic regression | 0.529 ± 0.000 | 0.819 ± 0.000 |
| SVC | 0.406 ± 0.000 | 0.805 ± 0.000 |

the best of our knowledge, this approach is the first to combine standard non-differentiable DT experts with MoE approach. Furthermore, we used MoËT in RL setting by mimicking DRL agents, in this way constructing RL policies that can be verified and are more interpretable than the DRL agents themselves. We showed a procedure to translate MoËT policies into SMT logic providing rich means for verification, and showed that MoËT models perform better than the previous state-of-the-art approach Viper and that they are also useful in the supervised regime.

**ACKNOWLEDGMENTS.** This work was supported by NSF grant CCF-1718903 to SK.

## References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, J. T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Briefings in bioinformatics 19 (6) (2018) 1236–1246.

[3] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature medicine 25 (1) (2019) 24–29.

[4] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, C.-M. Chen, Computer-aided diagnosis with deep

learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans, Scientific reports 6 (1) (2016) 1–13.

[5] M. Cicero, A. Bilbily, E. Colak, T. Dowdell, B. Gray, K. Perampaladas, J. Barfett, Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs, Investigative radiology 52 (5) (2017) 281–287.

[6] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Medical image analysis 35 (2017) 303–312.

[7] S. Ross, G. Gordon, D. Bagnell, A reduction of imitation learning and structured prediction to no-regret online learning, in: Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 627–635.

[8] O. Bastani, Y. Pu, A. Solar-Lezama, Verifiable reinforcement learning via policy extraction, in: Advances in Neural Information Processing Systems, 2018, pp. 2499–2509.

[9] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, et al., Adaptive mixtures of local experts., Neural computation 3 (1) (1991) 79–87.

[10] M. I. Jordan, L. Xu, Convergence results for the EM approach to mixtures of experts architectures, Neural networks 8 (9) (1995) 1409–1431.

[11] S. E. Yuksel, J. N. Wilson, P. D. Gader, Twenty years of mixture of experts, IEEE transactions on neural networks and learning systems 23 (8) (2012) 1177–1193.

[12] A. Biere, M. Heule, H. van Maaren, T. Walsh (Eds.), Handbook of Satisfiability, Vol. 185 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2009.

[13] L. De Moura, N. Bjørner, Z3: An efficient SMT solver, in: International conference on Tools and Algorithms for the Construction and Analysis of Systems, Springer, 2008, pp. 337–340.

[14] P. Van Wesel, A. E. Goodloe, Challenges in the verification of reinforcement learning algorithms, Tech. rep. (2017).

[15] G. Amir, M. Schapira, G. Katz, Towards scalable verification of deep reinforcement learning, in: 2021 Formal Methods in Computer Aided Design (FMCAD), IEEE, 2021, pp. 193–203.

[16] H. Zhu, Z. Xiong, S. Magill, S. Jagannathan, An inductive synthesis framework for verifiable reinforcement learning, in: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, 2019, pp. 686–701.

[17] Y. Kazak, C. Barrett, G. Katz, M. Schapira, Verifying deep-rl-driven systems, in: Proceedings of the 2019 Workshop on Network Meets AI & ML, 2019, pp. 83–89.

[18] A. Verma, H. M. Le, Y. Yue, S. Chaudhuri, Imitation-projected programmatic reinforcement learning, arXiv preprint arXiv:1907.05431 (2019).

[19] X. Li, Z. Serlin, G. Yang, C. Belta, A formal methods approach to interpretable reinforcement learning for robotic planning, Science Robotics 4 (37) (2019).

[20] C. Niu, F. Wu, S. Tang, S. Ma, G. Chen, Toward verifiable and privacy preserving machine learning prediction, IEEE Transactions on Dependable and Secure Computing (2020).

[21] J. Wang, S. Pandit, Towards high-level, verifiable autonomous behaviors with temporal specifications, in: 2019 IEEE National Aerospace and Electronics Conference (NAECON), IEEE, 2019, pp. 92–99.

[22] E. Bacci, D. Parker, Verified probabilistic policies for deep reinforcement learning, arXiv preprint arXiv:2201.03698 (2022).

[23] J. Törnblom, S. Nadjm-Tehrani, Formal verification of input-output mappings of tree ensembles, Science of Computer Programming 194 (2020) 102450.

[24] J. Törnblom, S. Nadjm-Tehrani, Formal verification of random forests in safety-critical applications, in: International Workshop on Formal Techniques for Safety-Critical Systems, Springer, 2018, pp. 55–71.

[25] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 93.

[26] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[27] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, Ieee Access 8 (2020) 42200–42216.

[28] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowledge-Based Systems 214 (2021) 106685.

[29] E. Puiutta, E. Veith, Explainable reinforcement learning: A survey, in: International cross-domain conference for machine learning and knowledge extraction, Springer, 2020, pp. 77–95.

[30] L. Wells, T. Bednarz, Explainable ai and reinforcement learninga systematic review of current approaches and trends, Frontiers in artificial intelligence 4 (2021) 48.

[31] S. Shi, J. Li, G. Li, P. Pan, K. Liu, Xpm: An explainable deep reinforcement learning framework for portfolio management, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1661–1670.

[32] A. Ayala, F. Cruz, B. Fernandes, R. Dazeley, Explainable deep reinforcement learning using introspection in a non-episodic task, arXiv preprint arXiv:2108.08911 (2021).

[33] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: KDD, 2016.

[34] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, arXiv preprint arXiv:1805.10820 (2018).

[35] X. Niuniu, L. Yuxun, Notice of retraction: Review of decision trees, in: 2010 3rd international conference on computer science and information technology, Vol. 5, IEEE, 2010, pp. 105–109.

[36] S. B. Kotsiantis, Decision trees: a recent overview, Artificial Intelligence Review 39 (4) (2013) 261–283.

[37] O. Irsoy, O. T. Yıldız, E. Alpaydın, Soft decision trees, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE, 2012, pp. 1819–1822.

[38] W. Zhao, Y. Gao, S. A. Memon, B. Raj, R. Singh, Hierarchical Routing Mixture of Experts, arXiv preprint arXiv:1903.07756 (2019).

[39] T. Hester, P. Stone, Texplore: real-time sample-efficient reinforcement learning for robots, Machine learning 90 (3) (2013) 385–429.

[40] C. Bucilu, R. Caruana, A. Niculescu-Mizil, Model compression, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 535–541.

[41] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[42] J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey, International Journal of Computer Vision 129 (6) (2021) 1789–1819.

[43] Z. Wang, Y. Wei, F. Wu, Knowledge distillation based cooperative reinforcement learning for connectivity preservation in uav networks, in: 2021 International Conference on UK-China Emerging Technologies (UCET), IEEE, 2021, pp. 171–176.

[44] L. Breiman, N. Shang, Born again trees, University of California, Berkeley, Berkeley, CA, Technical Report 1 (1996) 2.

[45] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born again neural networks, arXiv preprint arXiv:1805.04770 (2018).

[46] A. A. Rusu, S. G. Colmenarejo, Ç. Gülçehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, R. Hadsell, Policy distillation, in: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.

[47] A. Koul, A. Fern, S. Greydanus, Learning finite state representations of recurrent policy networks, in: ICLR, 2019.

[48] A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, T. Furlanello, Learning causal state representations of partially observable environments, CoRR (2019).

[49] A. Tsantekidis, N. Passalis, A. Tefas, Diversity-driven knowledge distillation for financial trading using deep reinforcement learning, Neural Networks 140 (2021) 193–202.

[50] Z. Gao, K. Xu, B. Ding, H. Wang, Knowru: Knowledge reuse via knowledge distillation in multi-agent reinforcement learning, Entropy 23 (8) (2021) 1043.

[51] P. Abbeel, A. Y. Ng, Apprenticeship learning via inverse reinforcement learning, in: ICML, 2004.

[52] S. Schaal, Is imitation learning the route to humanoid robots?, Trends in cognitive sciences (1999).

[53] Z. C. Lipton, The mythos of model interpretability, arXiv preprint arXiv:1606.03490 (2016).

[54] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529.

[55] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, N. De Freitas, Dueling network architectures for deep reinforcement learning, arXiv preprint arXiv:1511.06581 (2015).

[56] A. Frank, A. Asuncion, et al., Uci machine learning repository, 2010, URL http://archive.ics.uci.edu/ml 15 (2011) 22.

[57] R. Kohavi, Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid., in: Kdd, Vol. 96, 1996, pp. 202–207.

[58] A. G. Barto, R. S. Sutton, C. W. Anderson, Neuronlike adaptive elements that can solve difficult learning control problems, IEEE transactions on systems, man, and cybernetics (5) (1983) 834–846.

[59] R. S. Sutton, Generalization in reinforcement learning: Successful examples using sparse coarse coding, in: Advances in neural information processing systems, 1996, pp. 1038–1044.

[60] A. W. Moore, Efficient memory-based learning for robot control (1990).

[61] OpenAI Baselines, `https://github.com/openai/baselines`.

**Algorithm 2** Viper training [8]

**procedure** VIPER (MDP $e$, TEACHER $\pi_t$, Q-FUNCTION $Q^{\pi_t}$, ITERATIONS $N$)
  Initialize dataset and student: $D \leftarrow \emptyset, \pi_{s_0} \leftarrow \pi_t$
  **for** $i \leftarrow 1$ to $N$ **do**
    Sample trajectories and aggregate: $D \leftarrow D \cup \{(s, \pi_t(s)) \sim d^{\pi_{s_{i-1}}}(e)\}$
    Sample dataset using Q values: $D_s \leftarrow \{(s, a) \in I \sim D\}$
    Train decision tree: $\pi_{s_i} \leftarrow fit\_tree(D_s)$
  **return** Best policy $\pi_s \in \{\pi_{s_1}, ..., \pi_{s_N}\}$.

## Appendix A. Viper Algorithm

Viper algorithm is shown in Algorithm 2.

## Appendix B. Environments

In this section we provide a brief description of environments we used in our experiments. We used five environments from OpenAI Gym: CartPole, Acrobot, Mountaincar, Lunarlander, Pong and Pendulum.

### *Appendix B.1. CartPole*

This environment consists of a cart and a rigid pole hinged to the cart, based on the system presented by Barto et al. [58]. At the beginning pole is upright, and the goal is to prevent it from falling over. Cart is allowed to move horizontally within predefined bounds, and controller chooses to apply either *left* or *right* force to the cart. State is defined with four variables: $x$ (cart position), $\dot{x}$ (cart velocity), $\theta$ (pole angle), and $\dot{\theta}$ (pole angular velocity). Game is terminated when the absolute value of pole angle exceeds 12°, cart position is more than 2.4 units away from the center, or after 200 successful steps; whichever comes first. In each step reward of +1 is given, and the game is considered solved when the average reward is over 195 in over 100 consecutive trials.

### *Appendix B.2. Acrobot*

This environment is analogous to a gymnast swinging on a horizontal bar, and consists of a two links and two joins, where the joint between the

links is actuated. The environment is based on the system presented by Sutton [59]. Initially both links are pointing downwards, and the goal is to swing the end-point (feet) above the bar for at least the length of one link. The state consists of six variables, four variables consisting of sin and cos values of the joint angles, and two variables for angular velocities of the joints. The action is either applying *negative*, *neutral*, or *positive* torque on the joint. At each time step reward of $-1$ is received, and episode is terminated upon successful reaching the height, or after 200 steps, whichever comes first. Acrobot is an unsolved environment in that there is no reward limit under which is considered solved, but the goal is to achieve high reward.

*Appendix B.3. Mountaincar*

This environment consists of a car positioned between two hills, with a goal of reaching the hill in front of the car. The environment is based on the system presented by Moore [60]. Car can move in a one-dimensional track, but does not have enough power to reach the hill in one go, thus it needs to build momentum going back and forth to finally reach the hill. Controller can choose *left*, *right* or *neutral* action to apply left, right or no force to the car. State is defined by two variables, describing car position and car velocity. In each step reward of $-1$ is received, and episode is terminated upon reaching the hill, or after 200 steps, whichever comes first. The game is considered solved if average reward over 100 consecutive trials is no less than $-110$.

*Appendix B.4. Lunarlander*

This environment consists of a space ship and a landing pand, to which the ship should land. Controller can choose when to turn on the left engine, right engine or the main engine, thus controlling the movement of the ship. State is defined by: $x$ and $y$ coordinates of the lander, $v_x$ and $v_y$ velocities in the $x$ and $y$ direction, $\theta$ angle of the lander, $\alpha$ angular velocity, and two boolean values indicating if left or right leg is touching the ground. Episode finishes when lander crashes or comes to rest, after which it received appropriate reward. Firing main engine is $-0.3$ points, and each leg contact is 10 points. The game is considered solved if achieved reward is at least 200 points.

*Appendix B.5. Pong*

This is a classical Atari game of table tennis with two players. Minimum possible score is $-21$ and maximum is 21.

### *Appendix B.6. Pendulum*

The environment consists of a pendulum, and the goal is to swing it up so it stays upright. State is defined by: $\theta$—angle of the pendulum, and $\omega$—angular velocity of the pendulum. Note that the OpenAI gym environment instead of the state feature $\theta$ contains two features: $x$ (which is equal to $cos(\theta)$) and $y$ (which is equal to $sin(\theta)$). Action available is applying torque to the pendulum. In OpenAI gym action can take any value in range $[-2, 2]$. We discretize action space into 3 possible actions corresponding to torque of $-2$, 0, or 2. In each step reward obtained is equal to $-(\theta^2 + 0.1cdot\omega^2 + 0.001 \cdot torque^2)$. Thus, the maximum reward that can be obtained in a step is 0, which occurs when pendulum is upright, with zero velocity, and 0 torque is applied to the pendulum. Episode is of length 200.

## Appendix C. Model training parameters

### *Appendix C.1. DRL Agent Training*

In this section we present the architectures and hyperparameters used to train DRL agents for different environments.

For CartPole, we use policy gradient model as used in Viper. While we use the same model, we had to retrain it from scratch as the trained Viper agent was not available. We use 1 hidden layer with 8 neurons. We set discount factor to 0.99, number of epochs to $1,000$ and batch size to 50.

For Pong, we use a DQN network [54] model that is already trained (the same as used in Viper). This model originates from the OpenAI baselines [61].

For Acrobot, Mountaincar and Lunarlander, we implement our own version of dueling DQN network following [55]. We use 3 hidden layers with 15 neurons in each layer for Mountaincar, and 50 neurons in each layer for Acrobot and Lunarlander. We set the learning rate to 0.001, batch size to 30 in Mountaincar, 50 in Acrobot and Lunarlander, step size to $10,000$ and number of epochs to $80,000$ in Mountaincar, $50,000$ in Acrobot and Lunarlander. We checkpoint a model every $5,000$ steps and pick the best performing one in terms of achieved reward.

### *Appendix C.2. Viper and* MoËT *Training*

We used 40 iterations of DAgger, and $200,000$ as a maximum number of samples for training student policies. During evaluation, cummulative

reward is averaged across 100 runs in a given environment (250 in a case of CartPole).

We trained Viper for varying value of the tree maximum depth. The values used are: $[1, 15]$ in CartPole, $[1, 20]$ in Acrobot, $[1, 20]$ in Mountaincar, $[1, 30]$ In Lunarlander, and $[1, 35]$ in Pong.

We trained MoËT models for varying number of experts and their maximum depths. The number of experts used are: $[2, 8]$ in CartPole, $[2, 8] \cup [15, 16]$ in Acrobot, $[2, 8] \cup \{12, 16\}$ in Mountaincar, $[2, 8]$ in Lunarlander, and $\{2, 4, 8, 16, 32\}$ in Pong. The maximum depths of experts are: $[0, 7]$ in CartPole, $[0, 15]$ in Acrobot, $[0, 11]$ in Mountaincar, $[0, 20]$ in Lunarlander, and $[0, 29]$ in Pong. We used following learning rates for training MoËT models: $\{1, 0.3, 0.1, 0.01, 0.001, 0.0001, 0.00001\}$, while for the learning rate decay we used 1 (no decay) and 0.97 (learning rate is multiplied by this value after each epoch). As for the maximum number of epochs for MoËT training procedure we used values: $\{50, 100, 500\}$.

*Appendix C.3. Compute*

To run our experiments we used a cluster with nodes of the following configuration: Xeon CPU E5-2650 v3 (Haswell): 10 cores per socket (20 cores/node), 2.30GHz, 128 GB DDR4-2133. We used up to 10 such nodes when scheduling our experiments.

## Appendix D. SMT translation example

The CartPole $\text{MoË}\text{T}_\text{h}$ policy presented in Figure 6 is shown in Figure D.7. SMT formula that would encode the policy part (mapping input to a model decision) of CartPole verification formula would look as follows: `If(2.18cp + 7.22cv + 20.64pa + 25.33pv > -1, 1, 0)`. This $\text{MoË}\text{T}_\text{h}$ policy consists of the gating expressed by the inequality and two trivial expert decision trees of depth 0. Therefore, second and third part of the `If` formula are trivial. In case that decision trees were nontrivial, those parts of the formula would be expanded with nested if expressions.

A simple depth 2 Viper policy for CartPole is shown in Figure D.7. SMT formula that would encode the policy part of this formula would look like following: `If(pv < -0.033, If(pa < 0.039, 0, 1), If(pa < -0.037, 0, 1))`

The full formula for CartPole environment verification contains additional details, it is the conjunction of the formula encoding the policy, the safety

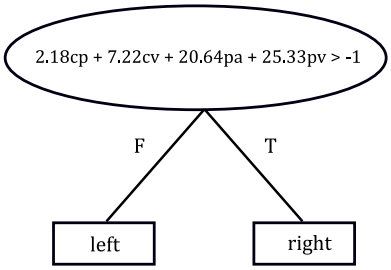


Figure D.7: Example CartPole MoËT$_h$ policy.

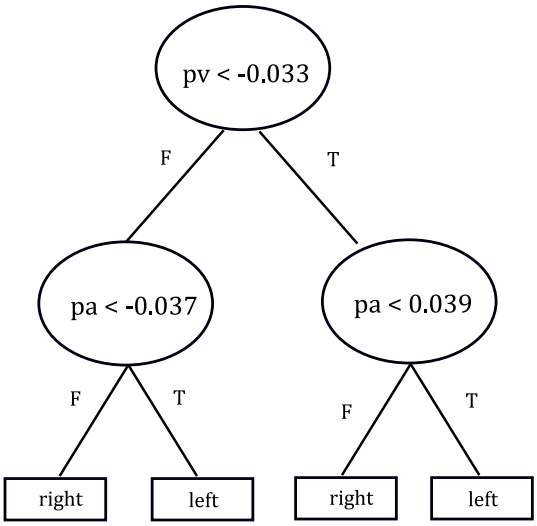


Figure D.8: Example CartPole Viper policy.

requirements and the environment dynamics, as illustrated by the formula in Section 5.

## Appendix E. Evaluation Results

Tables E.6, E.7, E.8, E.9, E.10, E.11 show data about models on the global Pareto front presented in Figure 3 of Section 6.

Tables E.12, E.13, E.14, E.15, E.16, E.17 show data about the models on the global Pareto after reevaluation is performed. This corresponds to data presented in Figure 4 of Section 6.

Table E.6: CartPole: global Pareto front data

| **Model** | **Configuration** | **Reward** | **Fidelity** |
|---|---|---|---|
| MoËT | E2-D0 | 200.00 | 0.998 |

Table E.7: Acrobot: global Pareto front data

| **Model** | **Configuration** | **Reward** | **Fidelity** |
|---|---|---|---|
| MoËT | E16-D11 | −72.12 | 0.936 |
| MoËT | E15-D11 | −71.95 | 0.936 |
| MoËT | E15-D11 | −71.81 | 0.921 |
| MoËT | E16-D9 | −71.67 | 0.921 |
| MoËT | E16-D0 | −69.83 | 0.916 |
| MoËT | E16-D0 | −68.68 | 0.907 |

Table E.8: Mountaincar: global Pareto front data

| **Model** | **Configuration** | **Reward** | **Fidelity** |
|---|---|---|---|
| MoËT$_{\text{h}}$ | E6-D9 | −107.00 | 0.984 |
| MoËT | E6-D7 | −106.83 | 0.984 |
| MoËT | E16-D7 | −105.90 | 0.983 |
| MoËT | E7-D8 | −104.28 | 0.982 |
| MoËT | E3-D7 | −103.86 | 0.979 |
| MoËT | E3-D10 | −103.82 | 0.977 |
| MoËT$_{\text{h}}$ | E3-D6 | −103.77 | 0.977 |
| MoËT | E7-D5 | −103.75 | 0.974 |
| MoËT | E3-D7 | −103.22 | 0.973 |
| Viper | D12 | −102.83 | 0.973 |
| MoËT | E2-D8 | −102.45 | 0.972 |
| Viper | D11 | −102.05 | 0.972 |
| MoËT$_{\text{h}}$ | E4-D4 | −101.40 | 0.971 |
| MoËT | E5-D5 | −101.09 | 0.966 |
| MoËT$_{\text{h}}$ | E8-D5 | −100.97 | 0.962 |
| MoËT$_{\text{h}}$ | E4-D5 | −100.96 | 0.961 |
| MoËT$_{\text{h}}$ | E2-D8 | −100.95 | 0.961 |
| MoËT$_{\text{h}}$ | E4-D5 | −98.85 | 0.960 |
| MoËT$_{\text{h}}$ | E4-D5 | −98.70 | 0.950 |
| MoËT | E4-D4 | −97.84 | 0.943 |
| Viper | D5 | −97.46 | 0.938 |
| MoËT | E7-D2 | −97.39 | 0.922 |
| MoËT | E4-D2 | −96.96 | 0.914 |
| MoËT$_{\text{h}}$ | E6-D1 | −96.78 | 0.912 |

Table E.9: Lunarlander: global Pareto front data

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E8-D17 | 204.13 | 0.792 |
| MoËT | E7-D17 | 210.79 | 0.767 |
| MoËT | E8-D17 | 217.33 | 0.765 |
| MoËT | E8-D17 | 225.24 | 0.755 |
| $\text{MoËT}_{\text{h}}$ | E8-D17 | 229.20 | 0.747 |
| MoËT | E6-D17 | 230.67 | 0.743 |
| $\text{MoËT}_{\text{h}}$ | E7-D0 | 239.96 | 0.666 |
| $\text{MoËT}_{\text{h}}$ | E7-D0 | 241.25 | 0.635 |
| MoËT | E6-D3 | 253.64 | 0.628 |
| $\text{MoËT}_{\text{h}}$ | E7-D0 | 261.86 | 0.547 |

Table E.10: Pong: global Pareto front data

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E16-D21 | 21.00 | 0.896 |

Table E.11: Pendulum: global Pareto front data

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E8-D16 | −170.00 | 0.988 |
| $\text{MoËT}_{\text{h}}$ | E7-D17 | −141.17 | 0.988 |
| MoËT | E4-D15 | −134.06 | 0.988 |
| MoËT | E6-D13 | −127.25 | 0.985 |
| $\text{MoËT}_{\text{h}}$ | E2-D12 | −120.31 | 0.979 |

Table E.12: CartPole: reevaluation Pareto

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E2-D0 | 200.00 | 0.998 |

Table E.13: Acrobot: reevaluation Pareto

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E15-D11 | −76.31 | 0.923 |
| MoËT | E15-D11 | −75.98 | 0.920 |
| MoËT | E16-D11 | −75.81 | 0.934 |
| MoËT | E16-D9 | −72.12 | 0.911 |
| MoËT | E16-D0 | −70.67 | 0.909 |
| MoËT | E16-D0 | −70.66 | 0.907 |

Table E.14: Mountaincar: reevaluation Pareto

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E3-D7 | −108.52 | 0.970 |
| MoËT | E7-D8 | −107.44 | 0.981 |
| MoËT | E16-D7 | −107.00 | 0.981 |
| MoËT | E3-D7 | −106.46 | 0.976 |
| MoËT | E3-D10 | −106.44 | 0.976 |
| MoËT | E6-D7 | −106.14 | 0.983 |
| $\text{MoËT}_\text{h}$ | E3-D6 | −106.09 | 0.973 |
| $\text{MoËT}_\text{h}$ | E6-D9 | −106.02 | 0.979 |
| Viper | D11 | −105.82 | 0.968 |
| MoËT | E2-D8 | −105.72 | 0.970 |
| Viper | D12 | −105.43 | 0.969 |
| MoËT | E7-D5 | −103.72 | 0.972 |
| $\text{MoËT}_\text{h}$ | E8-D5 | −102.92 | 0.958 |
| $\text{MoËT}_\text{h}$ | E2-D8 | −102.81 | 0.960 |
| MoËT | E5-D5 | −101.83 | 0.961 |
| $\text{MoËT}_\text{h}$ | E4-D5 | −101.75 | 0.960 |
| $\text{MoËT}_\text{h}$ | E4-D4 | −101.17 | 0.968 |
| $\text{MoËT}_\text{h}$ | E6-D1 | −99.82 | 0.906 |
| MoËT | E4-D2 | −99.47 | 0.910 |
| MoËT | E4-D4 | −99.37 | 0.936 |
| $\text{MoËT}_\text{h}$ | E4-D5 | −99.28 | 0.956 |
| MoËT | E7-D2 | −99.14 | 0.914 |
| Viper | D5 | −98.20 | 0.937 |
| $\text{MoËT}_\text{h}$ | E4-D5 | −97.88 | 0.950 |

Table E.15: Lunarlander: reevaluation Pareto

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E8-D17 | 178.93 | 0.762 |
| MoËT | E6-D17 | 180.40 | 0.751 |
| MoËT$_{\text{h}}$ | E8-D17 | 180.93 | 0.754 |
| MoËT | E8-D17 | 185.42 | 0.765 |
| MoËT | E7-D17 | 201.25 | 0.742 |
| MoËT | E8-D17 | 202.76 | 0.756 |
| MoËT$_{\text{h}}$ | E7-D0 | 232.45 | 0.660 |
| MoËT$_{\text{h}}$ | E7-D0 | 240.48 | 0.660 |
| MoËT$_{\text{h}}$ | E7-D0 | 247.97 | 0.537 |
| MoËT | E6-D3 | 256.90 | 0.588 |

Table E.16: Pong: reevaluation Pareto

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT | E16-D21 | 21.00 | 0.898 |

Table E.17: Pendulum: reevaluation Pareto

| Model | Configuration | Reward | Fidelity |
|---|---|---|---|
| MoËT$_{\text{h}}$ | E2-D12 | −177.01 | 0.976 |
| MoËT$_{\text{h}}$ | E7-D17 | −169.55 | 0.988 |
| MoËT | E4-D15 | −166.47 | 0.986 |
| MoËT | E6-D13 | −146.85 | 0.982 |
| MoËT | E8-D16 | −130.11 | 0.987 |