Research Statement

Abhinav Verma

Recent advances in machine learning have created the possibility of a Software 2.0 revolution, wherein code is generated based on the optimization of an evaluation criterion. However, a key requirement to real-world deployments of such software is generating learnt models that can be trusted by society. Current machine learning techniques rely heavily on Deep Neural Networks based models, which have significant fundamental drawbacks that make reliable learning difficult and learnt models susceptible to catastrophic failures. In some real world deployments of such models, bad outcomes have led to death and disability, thus eroding the public's trust in Artificial Intelligence (AI). The goal of my research is to generate trustworthy AI models, by integrating partial domain knowledge and experience based neural learning.

My research combines ideas from formal methods and machine learning to efficiently build models that are reliable, transparent, and secure. This means that such a system can be expected to learn desirable behaviors with limited data, while provably maintaining some essential correctness invariant and generating models whose decisions can be understood by humans. I believe that we can achieve these goals via Neurosymbolic learning, which establishes connections between the symbolic reasoning and inductive learning paradigms of artificial intelligence. My research is developing new theoretical foundations, algorithms, and tools in this area.

1 Prior Work

Current machine learning models are dominated by Deep Neural Networks (DNN), because they are capable of leveraging gradient-based algorithms to optimize a specific objective. However, neural models are considered "black-boxes" and are often considered untrustworthy due to the following drawbacks:

1. Hard to interpret: this makes these models hard to audit and debug.
2. Hard to formally verify: due to the lack of abstractions in neural models they are often too large to verify for desirable behavior using automated reasoning tools.
3. Unreliable: neural models have notoriously high levels of variability, to the extent that the random initialization of the weights can determine whether the learner finds a useful model.
4. Lack of domain awareness: neural models lack the ability to bias the learner with commonsense knowledge about the task or environment.

While these issues are well acknowledged in the ML community, most existing approaches tackle these problems individually and are unsuited for creating models that address all four drawbacks simultaneously. For example, existing interpretability tools do not provide a mechanism to make the network more amenable to formal verification. DNN verification techniques suffer scalability issues that reduce their applicability. Known regularization techniques often introduce a bias whose effects are hard to interpret or verify. And finally, domain awareness is sometimes implicitly encoded by pre-training on related tasks, but this pre-training is computationally expensive and has relatively few theoretical guarantees.

My current research focuses on addressing these four drawbacks simultaneously by automatically generating Neurosymbolic Models. These models combine learnt neural models with partial symbolic knowledge expressed via programs in a Domain Specific Language (DSL). At a high level, the neural model can perform learning via gradient-based methods and this information is then distilled into a constrained programmatic model. The constraints act as a mechanism to introduce symbolic domain knowledge into the learning process. The two models can be combined

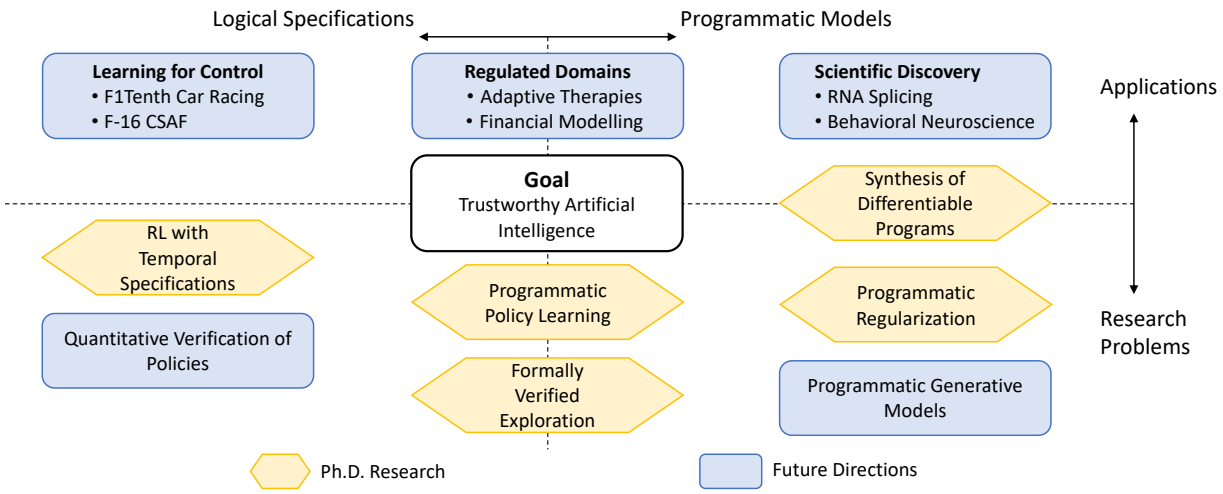

Fig. 1. Overview of my research around the theme of Trustworthy AI. Topics on the left and right leverage logical specifications and programmatic models, respectively. Foundational research problems are towards the bottom, and applications are towards the top.

in a variety of ways, which provide a technique to balance the relative benefits and drawbacks of each. In already published work, we have established that neurosymbolic models provide a principled mechanism to combat all four of the above shortcomings of DNN based models, and hence create a promising path towards trustworthy AI. An overview of my current and future research is given in Figure 1.

The intuition behind this work is that structured programs in a high level DSL have four key benefits. First, the DSL can be designed to be human-readable and is hence more interpretable than a DNN. Second, due to the availability of higher-level abstractions these models have parsimonious representations and are hence more amenable to formal verification techniques which can reason about the learned models and check consistency with desirable properties. Third, the DSL can be used to provide primitives that act as regularizers during learning, hence creating a more reliable model. Finally, the language can be used to encode commonsense knowledge thus providing a mechanism to programmatically modify the learner's inductive bias.

1.1 Programmatic Policies

We developed these ideas in the setting of synthesizing reinforcement learning policies, in the Programmatically Interpretable Reinforcement Learning (PIRL) framework. In summary, we place syntactic restrictions on the model via a user specified DSL. Our goal is to automatically find a program in this language, which maximizes the agent's expected aggregate reward in the environment. An example of this approach, is to synthesize programs that control a car's acceleration and steering to drive it around a track. Figure 2 shows the kind of high-level program our method finds for acceleration, when the DSL provides Propotional-Integral-Derivative (PID) controllers as primitives in the language. In general, the DSL is designed to provide high-level

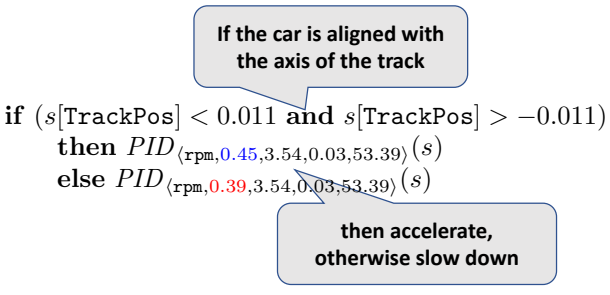

Fig. 2. An interpretable program for acceleration, automatically discovered in our framework.

abstractions that are known to be useful for the underlying domain. The automatically generated programs are hence parsimoniously represented in a structured programming language, which is similar to how a human expert would write such code.

A key challenge in PIRL is that the space of programs is typically vast and non-smooth making direct search intractable. Our approach to this question, an algorithm called Neurally Directed Program Search (NDPS), uses deep-RL to compute an initial approximation of the desired program, then uses this neural net as an "oracle" that guides program synthesis. This technique allows us to leverage recent advancements in quantitative program synthesis and SMT solvers.

1.2 Neurosymbolic Learning

As the NDPS algorithm requires a neural oracle, it can find performant policies only when deep-RL techniques are able to find performant policies independently. To address this shortcoming, we developed a novel meta-algorithm called Imitation-Projected Programmatic Reinforcement Learning (PROPEL), which is based on mirror descent, program synthesis, and imitation learning. The PROPEL framework leverages neurosymbolic learning to generate programmatic policies, by creating a neurosymbolic policy class which mixes neural and programmatic policy representations. This allows us to cast our learning task as optimization in a constrained policy space, and solve this problem using a "update-and-project" perspective that takes a gradient step into the unconstrained neurosymbolic space and then projects back onto the constrained programmatic space. Essentially, the PROPEL algorithm establishes a synergistic relationship between deep-RL and program synthesis, using synthesized programs to regularize deep-RL and using the gradients available to deep-RL to improve the quality of synthesized programs. This principled mechanism to create a neurosymbolic learner integrates symbolic knowledge with gradient-based optimization.

The domain knowledge embodied in the programming language acts as a form of regularization. This allows us to prove that neurosymbolic models can learn more reliably than traditional deep-RL methods. A thorough theoretical analysis of PROPEL characterizes the impact of approximate gradients and projections, providing promising expected regret bounds and finite-sample guarantees with the minimal assumption that the projection error is bounded. This analysis provides confidence that the PROPEL framework can be reliably applied to a variety of RL environments.

1.3 Logical Specifications

The parsimonious DSL constrained programs generated by these algorithms are both human interpretable and machine verifiable via off-the-shelf formal verification tools. In subsequent work we have shown that the PROPEL framework can be used to generate performant neurosymbolic controllers with provable safety guarantees, provided as logical constraints, which are not violated even during the learning process. This creates an avenue for the deployment of learning enabled controllers for safety-critical cyber-physical systems.

Specifications provided as temporal formulas can also be used to guide RL agents during training. This is especially useful in environments with sparse rewards, where the learner does not get a useful signal until a goal condition is met. Traditionally such problems have been solved by complicated reward engineering, which often needs to be tuned for every task and environment individually. Explicit planning is not possible in such cases, since the environment transitions are not known. Our approach uses a temporal specification to generate an abstract model, that is refined during the learning process, to guide the learner even when an explicit goal is not met. Our results establish that this technique can reliably find optimal policies for long-horizon tasks, that cannot be completed by exiting RL approaches without significant reward engineering.

2 Future Work

There has been tremendous interest in developing and applying neurosymbolic learning techniques to a variety of domains in recent years. However, it is clear that research in this area has only scratched the surface of the various possibilities, and there remains significant room for research in both effective DSL design and neurosymbolic learning algorithms.

One of my long-term research goals is to help lay the theoretical and infrastructural foundations required to make prototyping and deploying neurosymbolic learning as accessible as gradient based machine learning approaches. This goal is guided by the principle that reducing the learning curve of a new paradigm greatly increases its adoption and reach in both academia and industry. For example, the exponential rise of deep learning can be traced to the availability of libraries such as Torch (first developed in 2002 at IDIAP Research Institute), Theano (2007, University of Montreal) and others. My ongoing and future work is connected by the goal of building neurosymbolic pipelines that can be easily applied to a variety of domains and learning paradigms, and will hence spur the next generation of advancements in machine learning.

Concretely, such pipelines require two distinct components to be user friendly: creating domain specific primitives that have proven efficacy for certain tasks, and algorithmic innovations that are effective at overcoming the challenges of non-differentiable program learning in varied settings. This naturally decomposes the daunting task of creating this neurosymbolic infrastructure, into independent individual task-centric deployments that are promising research topics by themselves.

The work we discussed above has established the first set of domains and algorithms which deliver on the benefits of neurosymbolic learning. We have successfully used these techniques in both the supervised and reinforcement learning paradigms. Going forward, I will be working on extending and adapting these techniques to unsupervised learning. A promising avenue for this extension is via programmatic generative models, that can be used for data programming in domains that require controllable data generation which is not possible via black-box methods. There are also a multitude of deployment domains that can be explored with these techniques, and we discuss three broad categories below.

2.1 Learning for Control

Significant advances have been made in approaches that perform data-driven control, combining perspectives from machine learning and control theory. By providing a principled mechanism to guide the learner with partial symbolic domain knowledge, my work creates a promising bridge between traditional model-based design and controllers learnt via experience.

Control has been the underlying objective in many of the game simulators we have explored so far. Currently, I am developing a learning enabled collision avoidance system on DARPA's F-16 simulator, for a project on Assured Autonomy that will rely on the provable safety guarantees of such controllers. At UT's recently opened Robotics Center, I have also started deployments of neurosymbolic controllers on the F1Tenth car. Initial results have highlighted the robustness introduced by control primitives in the DSL, which help significantly reduce the performance degradation caused by the Sim2Real gap. For situations where providing qualitative guarantees is computationally infeasible, I am creating RL algorithms that integrate quantitative model checking techniques into the learning loop, to create controllers that probabilistically satisfy desirable properties.

2.2 Scientific Discovery

The availability of large amounts of data in almost every scientific field has led to machine learning playing an increasingly important role in scientific discovery. However, such techniques are seldom used for tasks other than pattern matching and clustering, in particular their role in devising hypotheses consistent with the data or imagining new experiments is significantly limited by the expressiveness of the models. Neurosymbolic models are capable of overcoming this deficiency.

In a recently accepted paper, we generate interpretable models for annotating animal behavior datasets. Apart from greatly reducing the burden of manual annotations, these expressive models can be used to guide further experimentation by focusing on factors that are identified by the model as most relevant for desirable behaviors. In ongoing work, I am working with biologists to generate models that reliably identify splicing points in a RNA sequence, and can describe the splicing mechanism for a given organism.

2.3 Regulated Domains

As AI based systems are included in more user facing applications, regulators are increasingly demanding clearer accountability for decision making processes. Notable examples of such regulatory requirements include the "right to explanation" clause in the European Union's General Data Protection Regulation (GDPR), and "plan of treatment" accountability under medical liability legislation in the US. Consequently, I have begun collaborating with researchers in healthcare and finance to generate interpretable and certifiable models for use in consumer facing applications.

Concretely, in healthcare I am working with Sepsis researchers to develop algorithms that generate interpretable adaptive therapies determined by a patient's statistics and their ongoing response to treatment. A key benefit of such a system is that it acts as a recommendation engine for attendant physicians and hence faces fewer regulatory requirements for deployment. In finance, I am working with JP Morgan Chase's AI research team to develop compliant models for financial markets. This work was the basis of my fellowship award from Chase, and has shown promising results for further exploration.

3 Conclusion

For many domains deep neural networks are the current state of the art machine learning method for generating performant models. My research aims to simultaneously address four fundamental drawbacks of such models; interpretability, verifiability, reliability and domain awareness. In published work we have formalized Programmatic Reinforcement Learning and provided empirical and theoretical evidence to show that we can generate models, without sacrificing performance, that do not suffer from these shortcomings.

These four drawbacks of neural models are fundamental impediments to the deployment of AI in real-world applications, as they create a deficiency of trust among regulators and the general public. In ongoing and future work I am researching methods to generate neurosymbolic models for a variety of new domains, and consequently laying the foundations for the wide scale adoption of neurosymbolic learning across academia and industry.

This work has created new connections between Machine Learning and Formal Methods, and has already helped create cross-disciplinary collaboration opportunities. By creating a principled mechanism to integrate symbolic knowledge with gradient based learning, this research program provides a promising path to Trustworthy Artificial Intelligence.