Neuro-Symbolic Model-based RL with Logical Neural Network

Don Joven Agravante[1] , **Daiki Kimura**[1] , **Michiaki Tatsubori**[1] ,
Asim Munawar[1] and **Alexander Gray**[1]
[1]IBM Research
don.joven.r.agravante@ibm.com, {daiki, mich}@jp.ibm.com, {asim, alexander.gray}@ibm.com

Abstract

The recent emergence of Neuro-Symbolic Agent (NeSA) approaches to natural language-based interactions calls for the investigation of model-based approaches. In contrast to model-free approaches, which existing NeSAs take, learning an explicit world model has an interesting potential especially in the explainability, which is one of the key selling points of NeSA. To learn useful world models, we leverage one of the recent neuro-symbolic architectures, Logical Neural Networks (LNN). Here, we describe a method that can learn neuro-symbolic world models on the TextWorld-Commonsense set of games. We then show how this can be improved further by taking inspiration from the concept of proprioception, but for conversation. This is done by enhancing the internal logic state with a memory of previous actions while also guiding future actions by augmenting the learned model with constraints based on this memory. This greatly improves the game-solving agents performance in a TextWorld setting, where the advantage over the baseline is an 85% average steps reduction and ×2.3 average score.

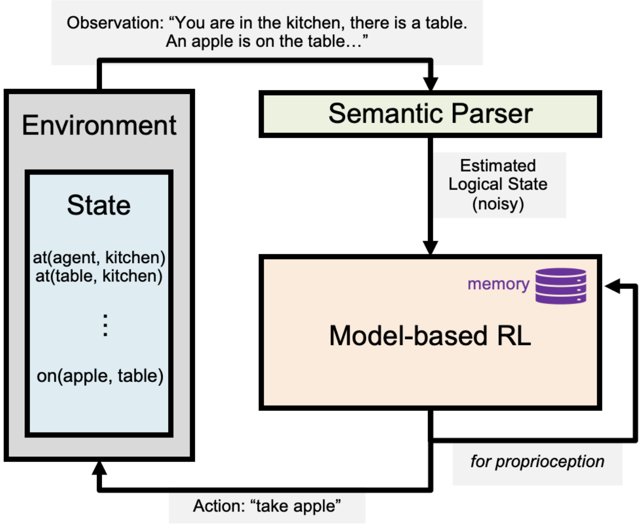

Figure 1: Simplified overview of the problem setting and agent. The observation from the environment describes the state in a natural language form. Our agent first uses a semantic parser to produce an estimate of the logical state. We then use this in a model-based RL module that learns a logical world model to plan actions to be taken. We augment the RL agent with a specific memory component inspired from proprioception.

1 Introduction

Recent emergence of neuro-symbolic (NS) approaches include natural language-based sequential decision making [Kimura *et al.*, 2021c; Chaudhury *et al.*, 2021; Kimura *et al.*, 2021a; Kimura *et al.*, 2021b]. They propose a model-free approach of learning a logical policy, and tested with interactive-text games [Narasimhan *et al.*, 2015; Côté *et al.*, 2018; Hausknecht *et al.*, 2020; Murugesan *et al.*, 2021], which have become an interesting benchmark in the intersection of natural language processing and sequential decision making. NS approaches give the direct explainability of what is learned and allow natural integration of external knowledge as logic. Despite that, existing NS approaches are of model-free reinforcement learning (RL) but it would be useful if we could have model-based approaches that are potentially more sample efficient and can reach higher cumulative rewards as shown by neural world models [Hafner *et al.*, 2019; **?**]. In contrast to these, a logical world model learned using NS approaches would allow an agent to use logical reasoning which enables us to obtain a trace of logical steps for better explainability. In fact, several sets of benchmarks and game environments have been proposed such as TextWorld [Côté *et al.*, 2018], Jericho [Hausknecht *et al.*, 2020] and TextWorld Commonsense (TWC) [Murugesan *et al.*, 2021], which are far too complicated to solve without reasoning and common sense, compared to the original game setting [Narasimhan *et al.*, 2015]. Also, existing implementations of NS agents do not start from natural language but instead use the logical facts provided from the game engines.

In this paper, we focus on the problem of learning logical world models in NS methods. The main research question to be addressed is then how we can learn such models for text-based games using a general semantic parser. As a state-of-the-art interactive-text agent, GATA [Adhikari *et al.*, 2020] constructs belief graphs used to enhance deep RL methods.

In contrast to understanding the world state in a latent space, we want to explicitly use the logical world models to plan optimal action sequences and to provide direct explainability of the decision making policy. For the explainability purpose, we leverage general semantic parsing, following one of the early work constructing knowledge graphs [Ammanabrolu and Riedl, 2019].

An overview of our proposed method is depicted by Figure 1. The left side depicts that the environment state can be sufficiently approximated as a set of logical facts. Continuing in the top right, the agent can get textual observations of the environment. We assume that we have a *semantic parser* [Drozdov *et al.*, 2022] that converts these observations into a logical form. In the real situation the semantic parsing is good, but won't be perfect, hence we require that our agent should be capable of handling noisy logical states. From such states, our agent should produce suitable actions for accomplishing its tasks in the environment.

The main contributions of this paper are: the proposal of a novel world model-learning method with a neuro-symbolic approach, and its experimental results with TWC.

2 Problem Definition

Text-based games are often modelled with the RL problem setting in mind as Partially Observable - Markov Decision Processes (PO-MDP) [Côté *et al.*, 2018; **?**]. As a first approach, we add an assumption - that the semantic parser can remove partial observability and that we are dealing with an MDP. At each time step the agent uses the information in a state, s, to take an action, a, which transitions the state to the new state, s' according to the state transition function T such that $s' = T(s, a)$. While acting in this environment the agent also gets rewards, r, according to an unknown reward function, R, such that $r = R(s, a)$. In the model-free RL setting, the agent learns a policy or value function which directly governs the actions. Here, we are interested in the model-based RL setting where the agent learns a model of the world which usually consists of both T and R. This model can then be used with planning methods to find the optimal actions.

Based on the classical model-based RL setting, our problem has two more important specifications. First, we assume that our environment is *relational*, similar to [Lang *et al.*, 2012]. This means that all actions and states are composed of relational logic. They may be in the propositional form but there must be a corresponding lifted form that has a consistent *meaning*. For example, the propositional state, *on(book,table)* can be abstracted or *lifted* into *on(x,y)* with predicate, on, and the variables, (x, y). The first assumption is that all states and actions handled by the agent are in this relational lifted form. This assumption can be handled as a design specification of the semantic parser. The second assumption is that the goal state is given. This is a weaker assumption that is already used in current RL research, the so-called *goal-conditioned RL*. Here, it allows us to concentrate only on learning T since R is no longer required for planning when we are given the goal state.

3 Learning Logical World Models

The problem of learning logical rules that explain a given set of logical examples can be cast into the general problem called Inductive Logic Programming (ILP) [Muggleton and De Raedt, 1994]. What needs to be done is then to cast our relational model-based RL problem into ILP form. But before going into that detail, it is important to note that relying on classical ILP has significant failings. In particular, it is not well suited to noisy data to the extent that a single erroneous data point may cause the whole system to fail.

However, newer methods that leverage neural networks have shown great promise on working even with noisy data [Evans and Grefenstette, 2018]. These are sometimes called neural ILP, differentiable ILP or neuro-symbolic ILP. These advances are the main impetus for us to research on the learning of logical world models.

We may use any such ILP method that is noise-resistant but here we use the Logical Neural Network (LNN) [Riegel *et al.*, 2020] as a Neuro-Symbolic AI framework. It is an end-to-end differentiable system that enables scalable gradient-based learning and it has a real-valued logic representation of each neuron that enables logical reasoning [Riegel *et al.*, 2020].

Action ILP with LNN

Now, getting back to the task of expressing our relational model-based RL problem as ILP, we first gather data samples which are triples of lifted logic, (s, a, s'). This is gathered by using an exploration policy to generate actions. This data collection may be done in an offline or online RL setting but we assume that a large enough *batch* is available in the online RL setting before we start the learning procedure. Here, we used a policy that uniformly randomly samples the action space but better exploration methods may be used, such as that outlined in [Lang *et al.*, 2012]. The improvement of better exploration is usually seen in *data efficiency* leading to faster convergence but using a sufficiently large amount of data won't change the benchmark scores on the TWC environment. We believe that a more rigorous treatment on the exploration of exponential but structured (logical) spaces merits its own research topic.

Given a batch of data samples, the learning procedure must produce an estimate of T. This T will be the *hypothesis* to be generated by our ILP. This is a set of logical rules that best fits the data. To make learning more efficient we need to narrow down the definition of T. Because we are ultimately interested in using T for planning, we define it as a set of planning operators where each one is a quadruple of $(\alpha, \beta, \gamma, \sigma)$. Each element is a set of logical conditions. The conditions (α, β) are pre-conditions where α are conditions that must be true for the action to be executable, β are ones that must be false. The conditions (γ, σ) are post-conditions where γ are ones made true by the action and σ are ones made false. These conditions are the lifted logic statements that comprise a state, s, and the set of all possible conditions is P.

We model each of the operator elements as an LNN conjunction operator whose inputs are P. The LNN learning procedure can learn weights for each of these inputs that correspond to real-valued logic [Riegel *et al.*, 2020; **?**]. For the LNNs of α and β, the inputs are given the corresponding logical values of the conditions in s. The output is true when action, a, corresponds and $s \neq s'$ otherwise it is false. For the LNNs of γ and σ, the inputs are given the logical values corresponding to the difference in the conditions of s and s' such that γ are the the conditions made true and σ those that are made false. The output is true when action, a, corresponds otherwise it is false.

Using these inputs and outputs to the LNN, gradient-based optimization can be used for supervised learning [Riegel *et al.*, 2020; **?**]. When learning converges, we have a set of weights for each of the corresponding elements. These may be interpreted as probabilistic transitions but here we simply threshold them and maintain a deterministic transition system for our final estimate of T. Given this operator transition model and the goal, we can be in any state and use classical planning methods to find a series of actions to reach the goal.

Conversational Proprioception as Memory-Based Constraints

To further enhance our logical model, we take inspiration from the concept of proprioception [Tuthill and Azim, 2018], which is the sensation of body position and movement critical to human experience, while it is typically absent from conscious perception. This concept is commonly used in imitation learning [Torabi *et al.*, 2019] and in robotics [Cong *et al.*, 2022]. In these domains, the type of sensors clearly distinguish the internal state measurement (proprioception) and external state measurement (perception). Combining both information sources is crucial to improving an agent's world model. We take inspiration from this to improve our logical world model estimate for text-based games or other tasks with logical state representations.

In general, proprioception is a prediction of the next state, $s' = \hat{T}(s, a)$, based on the existing knowledge of one's body dynamics in the form of the transition model estimate, $\hat{T}$, the current state, s, and the action taken, a. This additional information is crucial to help us disambiguate and better locate the next state. For our task where T is a logical model, we propose to augment our learned T with a set of proprioception rules, $\epsilon(s, a)$, such that our T will now be defined as $(\alpha, \beta, \gamma, \sigma, \epsilon(s, a))$. For our agent, we define ϵ very generally such that it only consists of 2 rules. First, it tracks state-action pairs that were already tried and augments the state with this information. This serves as a type of memory added onto the state. Second, it adds a precondition onto the transition models. This serves as a type of constraint on the actions. For our TWC agent, we defined preconditions that prevent state-action pairs from repeating. These 2 rules are general enough to apply to any TWC environment and possibly beyond to other conversational agents in general. We leave the design of further proprioception rules as a possible future work.

4 Experiment and Discussion

For evaluating the quality of the world model learned, we first qualitatively analyze the learned action models. Then we measure the interactive-text agent performance against a quantitative TextWorld benchmark. In this paper, we experiment on the TextWorld Commonsense (TWC) set of games [Murugesan *et al.*, 2021] with the same experimental settings.

Once we have a logical world model, we can use it with a planner. Here, we use the Fast-Downward systsem [Helmert, 2006]. For convenience, we convert the learned logical transition model into the PDDL (Planning Domain Definition Language) format by combining $(\alpha, \beta, \epsilon)$ into the preconditions and (γ, σ) into the effects. We also augment the state with $\epsilon(s, a)$.

Learned Models

We confirmed that the world models were meaningfully learned by any of model-based approaches. Figure 2 shows example learned action models in a converted PDDL form for an action *insert_into* (insert XX into YY) by model-based approaches from AMR-based logical facts. For our results, we first show some examples of the learned rules in our logical world model in Figure 2. Here, we can visually inspect the validity of the rules. For example with the left case, the effect would be that the object *v0* is at/in/on the container *v1* (*has_location-2*) but now it is no longer in the inventory (*carry-1*). This level of explainability is inherent in logical models although it requires careful inspection.

The effect of proprioception can be seen in the right-hand side of Figure 2. The predicate of *tried_insert_into* is from an AMR-based predicate *insert_into* but with the intention modality of the agent, which is encoded in first-order logic. This recognition of an already-performed action *insert_into* should contribute to avoiding repeatedly performing failed actions.

TWC Performance

It would be more interesting if we take altogether to see if the learned rules allow us to plan optimal actions in the world. To answer this, we present our results in Table 1. By running our complete framework, we can quantitatively compare against the benchmarks in [Murugesan *et al.*, 2021]. We can also see the effects of the important components of our agent.

Table 1 compares seven different methods/configurations corresponding to each row. The first row is a deep-learning-only method which is the best from the original benchmark in [Murugesan *et al.*, 2021]. The second and third rows are model-free neuro-symbolic methods. The fourth row is a planning result without any learning by using an ideal world model (assumed given) and access to (noiseless) logical game states. This serves as an upper bound for comparison only since having access the the ideal model and states is difficult or impossible in other applications. The fifth row is a model-based RL method given the ideal game-engine facts (equivalent to a perfect, noiseless semantic parser). The sixth row is our model-based RL method with a practical AMR-based semantic parser but without proprioception. Finally, the seventh row is our complete model-based RL method with a practical AMR-based semantic parser and proprioception module (memory-based constraints). In summary, the first four rows serve as comparison points and the last three rows shows the result of our method. Note that we have additional assumptions differing from the deep-learning-only setting of the original setup and we note these in the table as what type

	Semantic parsing	Handicap	Easy		Medium		Hard	
			Valid	Test	Valid	Test	Valid	Test
TWC agent (DL-only) [AAAI 2021]	Word embedding	(these are common) Admissible action Inventory Curated Common Sense	17.65 ± 3.62 85% ± 7%	18.00 ± 3.24 87% ± 5%	37.18 ± 4.86 72% ± 7%	43.08 ± 4.13 54% ± 17%	49.36 ± 7.50 46% ± 10%	49.96 ± 0.00 22% ± 0%
Model-free NeSA based on [EMNLP 2021]	Skipped	Game-engine facts	-	15.00 100%	-	28.60 100%	-	-
Model-free NeSA (REINFORCE)	AMR-based facts	-	-	32.28 ± 3.24 63% ± 5%	-	43.68 ± 5.36 38% ± 25%	-	49.48 ± 1.04 28% ± 13%
Planning (Model-based NeSA)	Skipped	Action transition Game-engine facts	2.4 100%	2.4 100%	4.4 100%	3.6 100%	13.6 100%	14.0 100%
Model-based NeSA (Learned action transition)	Skipped	Game-engine facts	2.4 ± 0.0 100%	2.4 ± 0.0 100%	4.4 ± 0.0 100%	3.6 ± 0.0 100%	13.6 ± 0.0 100%	28.4 ± 0.0 60.6%
Model-based NeSA	AMR-based facts	-	21.4 ± 0.0 57.1%	21.2 ± 0.0 42.9%	31.6 ± 0.0 38.5%	31.6 ± 0.0 50.0%	42.8 ± 0.0 20.6%	42.8 ± 0.0 24.2%
Model-based NeSA w/ proprioception	AMR-based facts	-	3.6 ± 0.0 100%	**4.0 ± 0.0** **100%**	7.6 ± 0.0 100%	**5.6 ± 0.0** **100%**	33.2 ± 0.0 64.7%	42.8 ± 0.0 24.2%

Table 1: Scores on the TextWorld Commonsense (TWC) set of games. Each cell contains 2 rows of metrics, each showing the average and standard deviations reported when available. The top row indicates the number of steps taken by an agent to finish the game. Lower is better because we want to reach the goal as fast as possible. The environment only allows a maximum of 50 steps. The bottom row is a normalized game score which is computed by taking the rewards obtained by the agent divided by the maximum rewards obtainable. Higher is better, where 100% means that the agent solved all the games in the benchmark set.

of semantic parsing and handicap are used. The TWC games are categorized into Easy-Medium-Hard with a validation and testing set for each as shown in the columns.

Comparing the result of our full method (last row) against current methods (first 3 rows) shows a significant improvement across the board. This shows the strength of the model-based NeSA framework against purely deep learning methods or the previously published model-free NeSA.

To see the effect of each component we can compare the results of the last four rows. Comparing the model-based NeSA with ideal semantic parsing (third to last row) against the planning upper bound, we can see that we can perfectly solve all except the test set of the hard games. After investigating, we found an interesting limitation wherein novel predicates appear in the test set that do not appear in any of the training or validation set. This is a current limitation of our system. Since we do not do any online learning during the test phase, there is no way to take these novel predicates into account. The significant effects of AMR-originated noise or lack of information can be seen by comparing the second-to-last and third-to-last rows. Here we see a significant degradation across metrics and datasets. However, the performance is still comparable or often better than the deep-learning-only benchmark of the first row. Comparing the last row (full method with proprioception module) and second-to-last row shows that we can recover most of the performance. We can also see that the metrics are competitive to the model-based approach from game-engine provided logical facts (3rd-last row). This shows the effectiveness of adding the proprioception module comprising both the memory and memory-based constraints.

5 Conclusion

```
(:action insert_into
:parameters (?v0 ?v1)
:precondition
(and
   (carry-1 ?v0)
   (not (has_location-2 ?v0 ?v1)))
:effect
(and
   (has_location-2 ?v0 ?v1)
   (not (carry-1 ?v0))))
```

```
(:action insert_into
:parameters (?v0 ?v1)
:precondition
(and
   (not (tried_insert_into ?v0 ?v1))
   (carry-1 ?v0)
   (not (has_location-2 ?v0 ?v1)))
:effect
(and
   (has_location-2 ?v0 ?v1)
   (not (carry-1 ?v0))
   (tried_insert_into ?v0 ?v1)))
```

Figure 2: Examples of the learned action models

We proposed a model-based RL agent for text-based games which comprises of a semantic parser producing logical states, a neuro-symbolic ILP module for learning logical world models, and an off-the-shelf planning system to produce optimal actions in the game world. We augment this with a proprioception-inspired module comprising both the memory and memory-based constraints. Our results and experiments show that each of the components are essential and our model-based NeSA agent outperforms previous benchmarks on the TextWorld Commonsense set of games.

Ethical Statement

There are no ethical issues.

References

[Adhikari *et al.*, 2020] Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. Learning dynamic belief graphs to generalize on text-based games. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3045–3057. Curran Associates, Inc., 2020.

[Ammanabrolu and Riedl, 2019] Prithviraj Ammanabrolu and Mark Riedl. Playing text-adventure games with graph-based deep reinforcement learning. In *Proceedings* *of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3557–3565, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Chaudhury *et al.*, 2021] Subhajit Chaudhury, Prithviraj Sen, Masaki Ono, Daiki Kimura, Michiaki Tatsubori, and Asim Munawar. Neuro-symbolic approaches for text-based policy learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 3073–3078, November 2021.

[Cong *et al.*, 2022] Lin Cong, Hongzhuo Liang, Philipp Ruppel, Yunlei Shi, Michael Görner, Norman Hendrich, and Jianwei Zhang. Reinforcement learning with vision-proprioception model for robot planar pushing. *Frontiers Neurorobotics*, 16:829437, 2022.

[Côté *et al.*, 2018] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532, 2018.

[Drozdov *et al.*, 2022] Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramón Fernandez Astudillo. Inducing and using alignments for transition-based AMR parsing. *CoRR*, abs/2205.01464, 2022. to appear in NAACL-22.

[Evans and Grefenstette, 2018] Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

[Hafner *et al.*, 2019] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[Hausknecht *et al.*, 2020] Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7903–7910, Apr. 2020.

[Helmert, 2006] Malte Helmert. The fast downward planning system. *Journal of Artificial Intelligence Research*, 26:191–246, 2006.

[Kimura *et al.*, 2021a] Daiki Kimura, Subhajit Chaudhury, Masaki Ono, Michiaki Tatsubori, Don Joven Agravante, Asim Munawar, Akifumi Wachi, Ryosuke Kohita, and Alexander Gray. LOA: Logical optimal actions for text-based interaction games. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 227–231, August 2021.

[Kimura *et al.*, 2021b] Daiki Kimura, Subhajit Chaudhury, Akifumi Wachi, Ryosuke Kohita, Asim Munawar, Michiaki Tatsubori, and Alexander Gray. Reinforcement learning with external knowledge by using logical neural networks. *arXiv preprint arXiv:2103.02363*, 2021.

[Kimura *et al.*, 2021c] Daiki Kimura, Masaki Ono, Subhajit Chaudhury, Ryosuke Kohita, Akifumi Wachi, Don Joven Agravante, Michiaki Tatsubori, Asim Munawar, and Alexander Gray. Neuro-symbolic reinforcement learning with first-order logic. In *Conference on Empirical Methods in Natural Language Processing*, pages 3505–3511, November 2021.

[Lang *et al.*, 2012] Tobias Lang, Marc Toussaint, and Kristian Kersting. Exploration in relational domains for model-based reinforcement learning. *Journal of Machine Learning Research*, 13(119):3725–3768, 2012.

[Muggleton and De Raedt, 1994] Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

[Murugesan *et al.*, 2021] Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Pushkar Shukla, Sadhana Kumaravel, Gerald Tesauro, Kartik Talamadupula, Mrinmaya Sachan, and Murray Campbell. Text-based rl agents with commonsense knowledge: New challenges, environments and baselines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9018–9027, May 2021.

[Narasimhan *et al.*, 2015] Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. Language understanding for text-based games using deep reinforcement learning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1–11, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[Riegel *et al.*, 2020] Ryan Riegel, Alexander G. Gray, Francois P. S. Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh K. Srivastava. Logical neural networks. *CoRR*, abs/2006.13155, 2020.

[Torabi *et al.*, 2019] Faraz Torabi, Garrett Warnell, and Peter Stone. Imitation learning from video by leveraging proprioception. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3585–3591. ijcai.org, 2019.

[Tuthill and Azim, 2018] John C. Tuthill and Eiman Azim. Primer: Proprioception. *Current Biology*, 28:R187–R207, 2018.