Action Grammars: A Cognitive Model for Learning Temporal Abstractions

Robert Tjarko Lange (robert.lange17@imperial.ac.uk)
Einstein Center for Neurosciences Berlin, Charitplatz 1, 10117, Berlin, Germany

Aldo Faisal (a.faisal@imperial.ac.uk)
Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

Abstract

Hierarchical Reinforcement Learning algorithms have successfully been applied to temporal credit assignment problems with sparse reward signals. State-of-the-art approaches require manual specification of sub-task structures, a sample inefficient exploration phase or lack semantic interpretability. Human infants, on the other hand, efficiently detect hierarchical sub-structures induced by their surroundings. In this work we propose a cognitive inspired Reinforcement Learning architecture which uses grammar induction to identify hierarchical sub-goal policies. More specifically, by treating an on-policy trajectory as a sentence sampled from the policy-conditioned language of the environment, we identify hierarchical constituents with the help of unsupervised grammatical inference. The resulting set of temporal abstractions is called *action grammar* (Pastra & Aloimonos, 2012) and can be used to accelerate and enable imitation, transfer and online learning.

Keywords: Decision Making; Reinforcement Learning; Computational Linguistics

Introduction

Human infants learn complex patterns in nature by observing role models and their behavior. Genetically inherited inductive biases allow to infer hierarchical rule-based structures from language, visual input as as well as auditory stimuli (M. C. Frank, Slemmer, Marcus, & Johnson, 2009; Marcus, Fernandes, & Johnson, 2007). Several MEG and fMRI studies provide evidence for a universal method of hierarchical language comprehension in the brain (S. L. Frank & Christiansen, 2018; Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Nelson et al., 2017) and a parallelism to motor control (Pastra & Aloimonos, 2012; Stout, Chaminade, Thomik, Apel, & Faisal, 2018). By processing trajectories of an expert, the infant is able to learn policies over higher level sequences of low level control elements. Inspired by these observations, this work proposes to overcome the problem of sub-structure discovery in Hierarchical Reinforcement Learning (HRL) by making use of grammatical inference. More specifically, the HRL agent uses grammar induction to extract hierarchical constituents from trajectory sentences. The proposed solution to the credit assignment problem is split into two alternating stages (see figure 1):

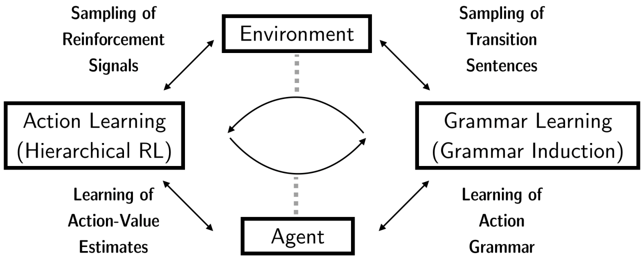


Figure 1: Online-inferred action grammars alternation loop.

1. **Grammar Learning**: Given episodic trajectories we treat the time-series of transitions as a sentence sampled from the language of the policy-conditioned environment. Using grammar induction (Nevill-Manning & Witten, 1997; Siyari, Dilkina, & Dovrolis, 2016) the agent extracts hierarchical constituents of the current policy. Based on the grammar estimate, temporally-extended actions are constructed which convey goal-driven syntactic meaning.
2. **Action Learning**: Using the grammar-augmented action space, the agent acquires new value information from interacting with the environment and refines his action-value estimates using Semi-Markov Decision Process (SMDP) Q-Learning (Bradtke & Duff, 1995). By operating at multiple time scales, the HRL agent is able to overcome difficulties in exploration and value information propagation. Afterwards, the agent samples simulated sentences by rolling out transitions from the improved policy.

By alternating between stages of grammar and action value learning the agent is able to iteratively reflect on their behavior in semi-supervised manner, to extract sub-components and to reduce the associated reward prediction error. Our experiments highlight the usability of the action grammars framework for imitation learning and transfer learning given an expert policy rollout. Furthermore, we derive promising results for an online version which iteratively refines grammar and value estimates.

Background

Temporal Abstractions. SMDPs extend Markov Decision Processes to incorporate not only reward and transition uncertainty but also time uncertainty. The time between individual decisions is modeled as a random variable, $\tau \in \mathbb{Z}_{++}$. It is characterized by the joint likelihood of transitioning from state $s \in \mathcal{S}$ to state s' in τ time steps given action m was pursued, $P(s', \tau | s, m)$. Thereby, SMDPs allow one to elegantly model the execution of actions which extend over multiple time-steps. A macro-action, $m \in \mathcal{M}$ specifies the sequential and deterministic execution of multiple (τ_m) primitive actions. Let $r_{\tau_m} = \sum_{i=1}^{\tau_m} \gamma^{i-1} r_{t+i}$ denote the accumulated and discounted reward for executing a macro. Value estimates can then be updated using SMDP-Q-Learning (Parr, 1998) in a model-free bootstrapping-based manner:

$$Q(s,m)_{k+1} = (1-\alpha) Q(s,m)_k + \alpha \left(r_{\tau_m} + \gamma^{\tau_m} \max_{m' \in \mathcal{M}} Q(s',m')_k \right)$$

In order to increase sample efficiency one can perform intra-macro updates for each state transition tuple $\{< s,a,r,s' >\}_{\tau_m}$ within the macro execution. The DQN (Mnih et al., 2015) objective can be adapted to the semi-Markov case:

$$L(\theta) := \mathbb{E}[(r_{\tau_m} + \gamma^{\tau_m} \max_{m' \in \mathcal{A} \cup \mathcal{M}} Q(s',m';\theta^-) - Q(s,m;\theta))^2]$$

Context-Free Grammars. Given a start symbol S, a formal grammar $(\Sigma, \mathbb{N}, S, \mathcal{P})$ produces an output of strings. Production rules $\mathcal{P}$ map a set of non-terminal vocabulary $\mathbb{N}$ either to another non-terminal or terminal string within the terminal vocabulary Σ. Context-free grammars (CFG) (Chomsky, 1959) constrain the set of productions to either map from one-to-one, one-to-none or one-to-many. A non-branching and loop-free CFG is called a straight-line grammar. Given a sample of sentences, grammar induction methods infer a grammar for a consistent language. Sequitur (Nevill-Manning & Witten, 1997) sequentially reads in all symbols and collects repeating subsequences of symbols into a production rule. Therewhile, the final encoded string is only allowed to have unique bigrams and production rules must be used more than once in the derivation of the string. In order to overcome Sequitur's problem of noise overfitting, k-Sequitur (Stout et al., 2018) has been proposed. Instead of replacing a bigram with a rule if the bigram occurs twice, it has to occur at least k times. As k increases the discovered CFG grammar becomes less sensitive to overfitting noise and the resulting grammar is more parsimonious in terms of inferred production rules.

Context-Free Action Grammars

Action sequences as well as communication convey goal-directed semantic meaning. They consist of hierarchical structures and are conditioned on the environment in which they are uttered in. Many real world problems require a hierarchy of subgoal achievements which increase in sequential difficulty and timescale. A trajectory obtained from traversing the current policy π can be viewed as a sample from the language generated by the grammar $L(\pi|E)$. Let the terminal vocabulary Σ consist of the primitive action space $\mathcal{A}$, hence $\Sigma = \mathcal{A}$. We denote $\vartheta^i \sim L(\pi|E)$ for $i = 1, \dots N_g$ trajectories. Given a set of trajectories, a CFG estimate $\hat{G}$ can be inferred and the resulting production rules can be transformed into macro-actions $\mathcal{M}^{\hat{G}}$ by recursively flattening the non-terminals. The action space of the action grammar HRL agent is then augmented such that $\mathcal{A}^{\hat{G}} = \mathcal{A} \cup \mathcal{M}^{\hat{G}}$. Depending on the generating policy of the compressed traces, one can construct several grammar-based HRL agents.

Expert & Transfer Grammars. If the traces ϑ^i are sampled from the language $L(\pi^\star|E)$ generated by the optimal policy, the agent can use the resulting grammar and flattened macros in an imitation learning setting. Before the onset of the first value learning stage, the action space is augmented with the optimal productions. Furthermore, an agent faced with learning a curriculum of tasks can make use of the optimal grammar of easier already solved tasks. Skills universal to all tasks do not have to be re-learned at every stage.

Online Inferred Grammars. If an episode successfully terminated, the grammar inference process identifies repeating sub-goal achieving patterns. We hypothesize that extracting action grammar sub-sequences compresses the temporal dimension of the credit assignment problem. After each grammar compression step, the action space is augmented with a new set of grammar macros. The previous set becomes inactive. In order to preserve value estimates between updates, we propose three solutions: (1) *Transfer learning* (Oquab, Bottou, Laptev, and Sivic (2014), see figure 2): In order to accommodate the variable set of grammar-inferred skills the size of the DQN output layer has to be updated with the newly macro-actions. Transferring the value-relevant features between action space augmentation, allows the agent to use the previously learned value characteristics. (2) *Grammar ER Buffer*: It is necessary to maintain a separate buffer system in order to store transition tuples specific to inactive previously inferred macro actions. At any given point the agent can only sample macro transitions which are associated with the currently active set of grammar macros. (3) *Intra-Macro Updates*: During the execution of a macro-action, one stores the overall macro transition tuple $< s_t, m_t, r_{t+\tau_m}, s_{t+\tau_m+1}, \tau_m, "on" >$ as well as the individual transitions $\{< s_i, a_i, r_i, s_{i+1}, 1, "on" >\}_{i=t}^{t+\tau_m}$. Thereby the agent is able to exploit all gathered transition experiences throughout the overall learning process.

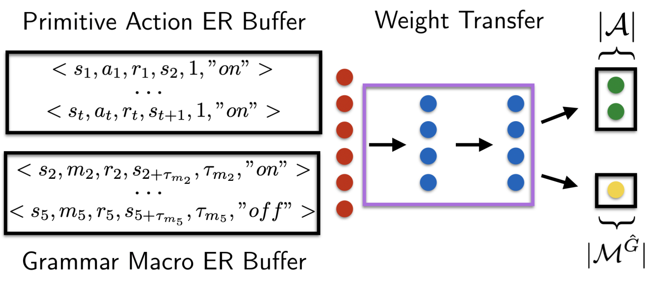

Primitive Action ER Buffer
$< s_1, a_1, r_1, s_2, 1, "on" >$
...
$< s_t, a_t, r_t, s_{t+1}, 1, "on" >$
$< s_2, m_2, r_2, s_{2+\tau_{m_2}}, \tau_{m_2}, "on" >$
...
$< s_5, m_5, r_5, s_{5+\tau_{m_5}}, \tau_{m_5}, "off" >$
Grammar Macro ER Buffer
Weight Transfer
$|\mathcal{A}|$
$|\mathcal{M}^{\hat{G}}|$

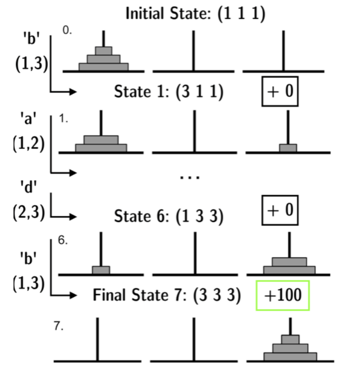
Figure 2: **Left.** Grammar Macro Experience Replay Buffer **Right.** Grammar Transfer DQN with adaptive output head.

The length of the sampled trace is going to increase or decrease over the course of the learning procedure. The regularization parameter of the k-Sequitur grammar inference algorithm has to be adapted accordingly. We experiment with a linear decaying scheme in order to assure that an appropriate amount of productions are inferred.

Experiments

The goal of the following experiments is to answer the following questions: (1) Does a grammar learned from optimal policy rollouts allow for rapid imitation learning? (2) Can CFG grammars be used in order to enhance curriculum learning by the means of transferring previously learned action grammars? (3) Is online grammar inference and action space adaptation able to structure the exploration process of the HRL agent? In order to illustrate our results we choose the general N-disk Towers of Hanoi (ToH) environment (see fig. 3) as well as a hierarchically structured gridworld task (see fig. 4).

Solving the N-disk ToH problem requires the agent to identify a hierarchical and recursive principle. By recursively moving $n-1$ disks onto an auxiliary pole and the n-th disk onto the target pole, the agent is able solve the sparse reward problem. Since such a routine can easily be formulated within a grammar parse tree, we hypothesize that the action grammars framework might provide an efficient solution.



Trace (ϑ): bafbcdbafec fbafbcdbcfecdbafbcdb

ϑ^{enc}	BCDfBEfDdBb
$\frac{\lvert\vartheta^{enc}\rvert}{\lvert\vartheta\rvert}$	0.355

$\mathbb{N}$	PR	$\mathcal{M}^{2-Seq}$
B	CEd	bafbcd
C	-	baf
D	-	ec
E	-	bc

Figure 3: **Left.** RL Formulation of the ToH Problem. **Right.** 2-Sequitur ToH (5 disks) Grammar-Macros.

The gridworld environment, on the other hand, provides a non-sparse reward design. The agent has to avoid poisonous items and moving blocks as well as collect food. The solution requires to solve large set of individually smaller subtasks.

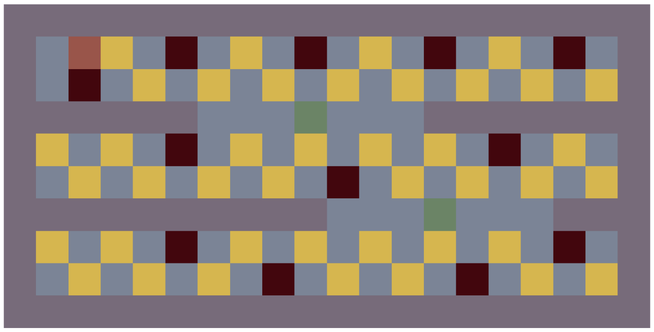


Figure 4: Hierarchically-Structured Grid World Environment.

Learning with Expert & Transfer Grammars. The right-hand side of figure 3 shows the grammar and resulting macros inferred from a trace of the optimal policy 5-disk ToH problem using the 2-Sequitur. The flattened production rule $B \rightarrow CEd \rightarrow bafbcd$ captures the recursive nature learned by the grammar. $C \rightarrow baf$ moves two disks on the auxiliary pole, while $E \rightarrow bc$ moves a third disk from source to target pole and one disk back onto the source pole. The HRL agent's action space is then augmented in the following way:

$$\mathcal{A}^{\hat{G}} = \mathcal{A} \cup \mathcal{M}^{2-Seq} = \mathcal{A} \cup \{bafbcd, baf, ec, bc\}$$

The left-hand side of figure 5displays learning results for different SMDP-Q-Learning agents with macro-actions defined by the production rules inferred from a single trace of the optimal policy. The grammar macros accelerate the learning progress of the agent. Furthermore, the variance of rollouts is reduced due to an increased robustness related to the temporal compression of the sequential problem. The right-hand side, on the other hand, assesses how well the grammars learned for a more simplistic environment (4 disks) generalizes to a more complex setting (5 disks). Again, the HRL agent is able to exploit the general grammar innate to the ToH problem.

Rob: Add nice figure with large fontsize a per updat more vari tion

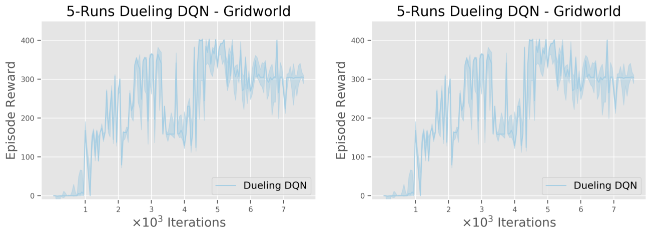


Figure 5: Expert & Transfer Grammar Experiments: **Left.** Expert Grammar. **Right.** Transfer Grammar. Averaged over 5 random seed. Median, 10th and 90th percentile.

Learning with Online Inferred Grammars. Figure 6 displays the results of the online grammar macro inference framework for the 4 disks ToH problem as well as the gridworld task.

Rob: Add ure and d scribe res

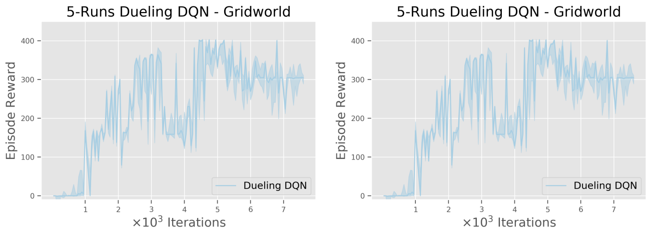


Figure 6: Online Grammar Experiments: **Left.** 4 Disk ToH Environment. **Right.** Gridworld Grammar DQN. Averaged over 5 random seed. Median, 10th and 90th percentile.

Conclusion

We have derived multiple algorithmic approaches which exploit powerful grammatical inference frameworks to identify temporally-extended actions. Our preliminary contributions are the following: (1) The CFG-based HRL agents are able to provide efficient and interpretable solutions to imitation and transfer learning tasks. (2) Alternating between grammar updates and learning action values is an effective method to learn of an optimal grammar as well as an optimal policy.

In future work we are interested in exploring stochastic grammars as well as their incorporation into model-based approaches. Furthermore, we intend to further analyze the relationship between grammar inference hyperparameters as well as exploration in HRL. Ultimately, we envision a dictionary of action which provides an expandable library of skills for agents which act in diverse naturalistic environments. This could provide a mayor contribution to a key endeavor in general artificial intelligence: Life-long learning.

: Add
sage on
ing and
avioral
notyping
omputa-
l psychi-
?

References

Bradtke, S. J., & Duff, M. O. (1995). Reinforcement learning methods for continuous-time markov decision problems. In *Advances in neural information processing systems* (pp. 393–400).

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, *157*, 81–94.

Chomsky, N. (1959). A note on phrase structure grammars. *Information and control*, *2*(4), 393–395.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental science*, *12*(4), 504–509.

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*, *0*(0), 1-6.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological science*, *18*(5), 387–391.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., . . . others (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 201701590.

Nevill-Manning, C. G., & Witten, I. H. (1997). Identifying hierarchical structure in sequences: A linear-time algorithm. *CoRR*.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1717–1724).

Parr, R. E. (1998). *Hierarchical control and learning for markov decision processes*. University of California, Berkeley Berkeley, CA.

Pastra, K., & Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1585), 103–117.

Siyari, P., Dilkina, B., & Dovrolis, C. (2016). Lexis: An optimization framework for discovering the hierarchical structure of sequential data. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1185–1194).

Stout, D., Chaminade, T., Thomik, A., Apel, J., & Faisal, A. A. (2018). Grammars of action in human behavior and evolution. *bioRxiv*, 281543.