Combining Structural Priors and Representation Learning for Operating Robotics Systems under Sparse Information

Artem Molchanov
University of Southern California

1 INTRODUCTION

The last decade has brought us a lot of innovations and breakthroughs in different areas of artificial intelligence (AI). We have seen machines outperforming humans on tasks like image recognition [4, 11, 16] and games [19, 22]. *Neural networks* (NN) are one of the driving forces in this wave of progress. As a result, successful applications of NN created a lot of excitement and, most importantly, they gave us a hope that we could solve a lot of problems in AI by simply leveraging large amounts of well annotated data with ground truth labels, target measurements and perfect demonstrations that, in this thesis, we will aggregate under the term *direct information*. In robotics we do not have access to such abundant sources of direct information because of several factors: robots are diverse and expensive to run and operate, both in terms of time and cost. Furthermore, robots often use very specific information, for example, motor commands or special properties of objects, such as orientation, friction coefficients of surfaces and materials or grasping points and configurations. These factors are highly dependent on specific robot hardware and conditions of experiments making standardization of datasets problematic. On top of that, in order to operate in the real world robots have to deal with a wide range of diverse conditions and adapt to new situations under noisy, incomplete and very limited data samples with non-trivial mapping to direct information. By analogy, we will aggregate data with such properties under the term *indirect information*. In other words, we can only assume *sparse* access to direct information available to the robot during its operation.

By *sparsity of information* we will refer to general lack of direct information or data required to infer direct information (i.e. information relevant for the task). Examples in robotics include availability of a just a few data samples with direct information or complete lack of them; availability of just a few data samples with good observability of the state (or other forms of direct information, for example, optimal actions) among many others with low observability; spatial and temporal sparsity of measurements, where data samples are available with very low spatial or temporal resolution or simply cannot cover the desired space. The central theme of this thesis are methods that can better leverage limited or sparsely available direct information and efficiently utilize sources of indirect information.

To pursue our goal, we would like to leverage one of the main advantages of neural networks - their representation (or expressive) power that allows us learning complex dependencies with low interpretability. On the other hand, we would like to combine it with inspiration from compressed sensing that allows exact inference of complex signals from temporally and spatially sparse measurements, by exploiting sparsity of underlying structures in the signal. We will go broader and enhance neural networks and the learning algorithms with different forms of priors by structuring architectures and learning algorithms. Thus, we can exploit regularities in data for efficient learning from sparse information and we can utilize indirect information extensively.

In this thesis, we will investigate how sparsity of information exhibits itself in typical robotics sub-components, such as i) perception, ii) control, and iii) task specification. We show how we can use deep neural networks in combination with special structures introduced into the learning pipeline to enable efficient adaptation under sparsity for successful operation of robotics systems in the context of different problems.

In the first part, we will show how we can overcome sparsity of perception of external contacts due to absence of measurements from more traditional sensors by utilizing complex and multi-modal indirect measurements from tactile sensors. In the second part, we will take a look at the problem of sparsity of rewards in reinforcement learning (RL) that relate to the sparsity of information for task interpretation. In the third part, we will take a look at the problem of safe and efficient learning of dynamics models and control policies with sparsity of failure data samples.

2 SPARSITY IN PERCEPTION

We start with the sparsity in perception. One aspect that separates robotics from other areas of application of AI is the necessity of operating in a physical world and deal with contacts. Thus, estimating different properties of contacts, such as locations and orientations of applied forces (later summarized under the term point-of-contact localization) is crucial for many areas of robotics. One of the most common use cases is tool manipulation, where contact localization between the tool and the environment (later referred as the external contact) enables task verification. For example, if a robot manipulates an object that it has to place on a table, it must detect that the contact force is perpendicular to the bottom of an object and originates on its bottom and not on the side.

Traditionally localization of external contacts is approached using either vision or force-torque sensors. Vision provides very sparse relevant observations for this problem, because the contact surface is usually occluded by the manipulated and surrounding objects in the environment. Force-torque sensors are not always available on each robot, due to high cost. They also drift over time and the localization methods based on force-torque sensors often require accurate calibration and knowledge of kinematics and dynamics models of the robot. Therefore, we can simply assume the scenario of ultimate sparsity of the information about external contacts, i.e. sparsity or absence of traditional sensory measurements for such estimation. Thus, there is a lot of value in developing alternative sources of information that might be available to the robot that could densify estimates from the traditional sources.

One potential source of such information are tactile sensors that remain underexplored for the problem of the external contact localization. Tactile sensors are still in their infancy. As a result, their application brings a lot of problems, such as hardware fragility, poor understanding of signals, lack of intuition for analysis, and absence of datasets. On the other hand, they gradually gain popularity, because of a strong evidence in neuroscience [10, 14] about the importance of tactile feedback for many manipulation related scenarios. Among others, BioTac [24] sensors have shown to contain a lot of relevant information for many contact related properties, which motivated us to leverage this sensor for our goal.

We use the representational power of neural networks to learn the complex and unintuitive mapping from BioTac signals to relative locations and orientations of external contact point in the wrist frame of the robot. We demonstrate that traditional framing of such problem as regression provides poor results due to a possible ambiguity of such estimation resulting in regression averaging. Therefore, we incorporate a more structured learning by discretizing the continuous space and proposing marginal classifiers (separate classifiers for each dimension) to mitigate the problem of the curse of dimensionality of the otherwise exponential number of labels.

Our proposed solution shows that such simple structuring allows us to significantly improve the estimation. We demonstrate that i) neural networks have the best performance in comparison to other state-of-the-art machine learning techniques, such as Gaussian processes (GP) and Spatio-Temporal Hierarchical Matching Pursuit (STHMP) [18]; and ii) we can reduce the localization error from 5cm to 2cm by structuring the problem as classification.

3 SPARSITY IN TASK SPECIFICATION

For a robot to do meaningful work in an environment and to adapt to constantly changing tasks, it is crucial to have a simple and robust way of communicating the desired task to the robot. However, this becomes increasingly harder with growing task complexity that the robot has to perform. In recent years, deep reinforcement learning (Deep RL) has enjoyed success in many different applications, including playing Atari games [19], controlling a humanoid robot to perform various manipulation tasks [2, 3], and beating the world champion in Go [22]. The success and wide range of use cases of RL algorithms is partly due to the very general description of the problem that RL aims to solve, i.e., to learn autonomous behaviors given a high-level specification of a task by interacting with the environment. Such high-level specification is provided by a reward function, which must be sufficiently descriptive as well as easy to optimize for an RL algorithm to learn efficiently. These requirements make the design of the reward function challenging in practice, creating a bottleneck for even a wider set of applications for RL algorithms. As a result, RL often faces the problem of task confusion and undesirable local minima, resulting in a wrong and sub-optimal behavior. To overcome these challenges the problem of designing a reward function has been tackled in various ways. They include: i) learning the reward function from human demonstrations in the field of inverse reinforcement learning (IRL) [1, 17], ii) initializing the reinforcement learning process with demonstrations in imitation learning [3, 15], and iii) creating reward shaping functions that aim to guide the RL process to high-reward regions [2, 21]. Even though all of these methods have shown promising solutions to some of the problems mentioned above, they present other significant challenges such as the requirement of domain expertise or access to demonstration data.

Ideally, one would like to learn from a simple sparse binary reward that indicates the completion of the task. Such a reward signal is natural for many goal-oriented tasks. It requires significantly less engineering effort, and in some cases can be used to learn very complicated skills from human feedback, where design of the reward function is very hard [5] or in situation when the robot has to be re-trained by nonprofessionals. Despite being attractive, this type of reward functions creates significant difficulties for learning, especially when data-greedy neural networks are used for learning of policy representations. This is due to the fact that it is very unlikely for an agent to generate the exact sequence of actions leading to solving the task from random exploration [6] resulting in sparse task specification and lack of feedback for policy learning.

Recent efforts focus on learning from such sparse reward signals by exploiting a structured approach to task learning via constructing a curriculum from a continuous set of tasks [9, 12]. The curriculum structure exploits the simple intuition that tasks initialized closer to the goal should be easier to solve. Proximity to the goal is defined either explicitly [12] or through the a chain of random actions needed to reach the state from the goal [9].

Nevertheless, all of these methods have a common disadvantage: they are designed for either single-start or single-goal scenarios. We address the situation in which the task contains both a continuous set of goals and a continuous set of initial conditions, thus broadening the applicability to a wide range of problems. To construct the curriculum we exploit an approach of designing the task learning in a form of cooperative game of two agents: the teacher and the student. The teacher is responsible for task proposals, i.e. it expands the growing region of states using random sampling, whereas the student learns state transitions with sparse rewards. Since the teacher-student setting resembles the idea of adversarial training of two agents, it encounters similar problems. One of these problems is balancing progress of the two agents during the training process. To overcome this problem, we apply the equilibrium principle through balancing the success ratio for the student. To implement this principle, we introduce a method of adaptively adjusting expansion of the growing region by controlling the teacher's sampling variance. Our variance adaptation algorithm is inspired by the integral control law from classic control theory.

Our results show that we can effectively learn in environments with continuous state, action spaces and multiple goals using sparse rewards. On top of that, our algorithm demonstrates adaptation to environments with diverse requirements for region growing expansion, which avoids tedious variance tuning.

4 SPARSITY IN DYNAMICS MODELS

Real world robotic systems face the problem of safe and efficient policy adaptation in the presence of unknown and changing dynamics. Safety is a strong limiting factor in acquiring data, such as states and actions corresponding to wrong behavior (e.g. crashes) that are essential for learning robust policies. In other words, safety requirements create sparsity of failure data samples (both in terms of observation/state space coverage and in terms of overall quantity) that are essential for learning undesirable types of behavior.

To attack this problem we will use our idea of leveraging representational power of neural networks to learn dynamics and policies of arbitrary complexity and we will cast two types of priors to exploit regularities of the problem. First, we would like to leverage recent advances in meta learning to create efficient priors on learned parameters for rapid adaptation under sparse (limited) data samples. Second, we will draw inspiration from the principle of sparse model identification to find a hierarchy of neural models with minimally needed complexity for efficient learning of dynamics under constrained computational resources of robotics systems.

For the first part, we would like to utilize dynamics randomization in simulation [20, 23, 25] and combine it with meta-learning [7, 8] for finding good initialization (prior) point in the parameter space of the model and the policy. Dynamics randomization is a concept of training a learner, such as a control policy, by constantly randomizing dynamical systems within a certain domain. It helps making the learner to be robust to dynamics perturbations, but may not allow the best performance for a particular dynamical system. Meta-learning (or learning to learn) is a family of algorithms generating algorithms that can efficiently adapt within a certain domain of tasks. Hence, this combination will enable i) learning a safe exploration policy for a wide range of dynamical systems, and ii) rapid adaptation of the target policy under just a few failure examples sampled from a target dynamical system.

In addition to meta learning we will add structured hierarchy of auxiliary controllers from simple to complex, where more complex models would compensate for significantly smaller portion of the control signal. This has a potential advantage of better sample efficiency of learning with separable models for different complexity and degrees of approximation (depending on the hardware resources available).

One interesting application of our approach is quick adaptation of quadrotor control policy for a particular quadrotor model. The main intuition of why such combination of techniques should help in this application domain comes from several factors. First, it has been already shown that it is possible to train a moderate stabilizing controller for a quadrotor using reinforcement learning in simulation and transfer it to a real system [13]. Their work assumes that the dynamics in simulation reasonably match the real world. In contrast, we seek for adaptation to completely unknown dynamics. Second, the modern meta learning approaches are only capable of adapting to tasks from the same or very similar domain. We see that variations of quadrotor dynamics falls into this category. Third, we can exploit the fact that for the sampling stage performed by the exploration policy the main objective is system identification. In other words, one does not need a good tracking controller for state sampling, since the main goal is acquisition of a decent state distribution in as safe manner as it is possible. Fourth, quadrotors retain a lot of necessary properties: they have relatively complex dynamics with underactuation and unstable states (meaning that applying null action leads to an inevitable crash), they can tolerate some limited number of crashes, they have constrained computational resources, and their state representation could be relatively low dimensional. Thus, this platform proposes significant challenge while retaining enough desired properties for inexpensive, but reliable evaluation of the proposed approach.

5 CONCLUSIONS

In this work, we investigate how combination of learning complex representations and structured priors can provide an effective solution for adaptation of robotics systems under sparsity of direct information. We investigate problems related to sparsity in different sub-components of various robotics systems within the context of specific robotics problems.

In the first part, we look at the problem of sparsity of information for point-of-contact estimation that originated from the absence of reliable measurements of other sensors. We propose to utilize tactile measurements as a novel source of information about external contact locations and orientations. We leverage the representational power of neural networks and propose disretization and marginalization of classifiers as a way of structuring the learning problem.

In the second part, we look at the problem of sparsity of task specification in the domain of reinforcement learning. Particularly, we investigate a problem of learning complex policies under sparse rewards. We use neural networks to represent the policy function and we structure the learning algorithm in a form of curriculum discovering cooperative game of two agents: the teacher agent and the student. To balance the progress of the two agents we adjust the teacher's variance of exploration by using inspiration from control theory.

In the third part, we look at the problem of sparsity of negative (failure) data samples for learning of complex control policies under unknown dynamics imposed by safety requirements. We investigate this problem in the context of end-to-end learning of a quadrotor's stabilizing controller with unknown dynamics model. We leverage expressive power of deep neural networks to represent the complex model dynamics and the policy. We use meta learning in combination with dynamics randomization in simulation to discover optimal simulated priors for fast parameter adaptation. To have flexible adaptation to computational constraints and further improve on sample efficiency of learning we propose a hierarchy of auxiliary error compensating models of increased representation complexity. Finally, we train a special sampling policy whose purpose is safe and efficient system identification to facilitate safety of exploration.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *ICML*. https://doi.org/10.1145/1015330.1015430

[2] Y. Chebotar, K. Hausman, et al. 2017. Combining Model-Based and Model-Free Updates for Trajectory-Centric Reinforcement Learning. In *ICML*. http://arxiv.org/abs/1703.03078

[3] Yevgen Chebotar, Mrinal Kalakrishnan, et al. 2017. Path integral guided policy search. In *ICRA*. 3381–3388. http://dblp.uni-trier.de/db/journals/corr/corr1610.html#ChebotarKYLSL16

[4] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. *CoRR* abs/1511.05960 (2015). http://arxiv.org/abs/1511.05960

[5] Paul F. Christiano, Jan Leike, et al. 2017. Deep Reinforcement Learning from Human Preferences. In *NIPS*. 4302–4310.

[6] Yan Duan, Xi Chen, et al. 2016. Benchmarking Deep Reinforcement Learning for Continuous Control. In *ICML*. 1329–1338.

[7] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL$ˆ2$: Fast Reinforcement Learning via Slow Reinforcement Learning. *CoRR* abs/1611.02779 (2016). arXiv:1611.02779 http://arxiv.org/abs/1611.02779

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, 1126–1135. http://proceedings.mlr.press/v70/finn17a.html

[9] Carlos Florensa, David Held, et al. 2017. Reverse Curriculum Generation for Reinforcement Learning. In *CoRL*. 482–495. http://arxiv.org/abs/1707.05300

[10] Antony W. Goodwin and Heather E. Wheat. 2004. Sensory Signals in Neural Populations Underlying Tactile Perception and Manipulation. *Annual Review of Neuroscience* 27 (2004), 53–77.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 1026–1034. https://doi.org/10.1109/ICCV.2015.123

[12] David Held, Xinyang Geng, et al. 2017. Automatic Goal Generation for Reinforcement Learning Agents. *CoRR* abs/1705.06366 (2017).

[13] Jemin Hwangbo, Inkyu Sa, Roland Siegwart, and Marco Hutter. 2017. Control of a Quadrotor With Reinforcement Learning. *IEEE Robotics and Automation Letters* 2, 4 (2017), 2096–2103. https://doi.org/10.1109/LRA.2017.2720851

[14] Roland S. Johansson and Randall J. Flanagan. 2009. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience* 10, 5 (April 2009), 345–359. https://doi.org/10.1038/nrn2621

[15] Mrinal Kalakrishnan, Ludovic Righetti, et al. 2012. Learning Force Control Policies for Compliant Robotic Manipulation. In *ICML*. http://icml.cc/2012/papers/642.pdf

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS 2012*. 1097–1105.

[17] Sergey Levine, Zoran Popovic, and Vladlen Koltun. 2011. Nonlinear Inverse Reinforcement Learning with Gaussian Processes. In *NIPS*. 19–27. http://papers.nips.cc/paper/4420-nonlinear-inverse-reinforcement-learning-with-gaussian-processes

[18] Marianna Madry, Liefeng Bo, Danica Kragic, and Dieter Fox. 2014. ST-HMP: Unsupervised Spatio-Temporal Feature Learning for Tactile Data. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[19] Volodymyr Mnih, Koray Kavukcuoglu, et al. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013).

[20] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2017. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. *CoRR* abs/1710.06537 (2017). arXiv:1710.06537 http://arxiv.org/abs/1710.06537

[21] Ivaylo Popov, Nicolas Heess, et al. 2017. Data-efficient Deep Reinforcement Learning for Dexterous Manipulation. *CoRR* abs/1704.03073 (2017). http://dblp.uni-trier.de/db/journals/corr/corr1704.html#PopovHLHBVLTER17

[22] David Silver, Aja Huang, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489. https://doi.org/10.1038/nature16961

[23] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*. IEEE, 23–30. https://doi.org/10.1109/IROS.2017.8202133

[24] N. Wettels, J.A. Fishel, and G.E. Loeb. 2014. Multimodal Tactile Sensor. In *The Human Hand as an Inspiration for Robot Hand Development*. Springer Tracts in Advanced Robotics, Vol. 95. Springer, 405–429.

[25] Wenhao Yu, Jie Tan, C. Karen Liu, and Greg Turk. 2017. Preparing for the Unknown: Learning a Universal Policy with Online System Identification. In *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017*, Nancy M. Amato, Siddhartha S. Srinivasa, Nora Ayanian, and Scott Kuindersma (Eds.). http://www.roboticsproceedings.org/rss13/p48.html