COMP 767 1–6

Diversity-Based Learning for Robotics

Andrew Holliday AHOLLID@CIM.MCGILL.CA
McGill ID: 260604560

Nikhil Rajiv Kakodkar NIKHIL.KAKODKAR@MAIL.MCGILL.CA
McGill ID: 260578689

Karim Koreitem KARIM.KOREITEM@MAIL.MCGILL.CA
McGill ID: 260460964

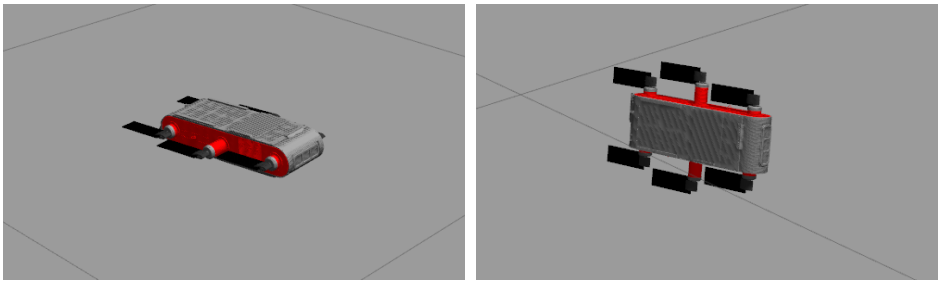
1. Introduction

In this work, we assess whether the diversity of state spaces accessible to a robot is a useful metric for measuring the benefits of adding more end-effectors to a given robot design. To that end, we attempt to apply an unsupervised learning technique for the learning of a diverse skill-set to a realistic robotic setting. We experiment with various configurations of the technique. Initial results are counter-intuitive, and suggest that the policy model is insufficient to the task at hand. Further experiments attempting to use the same policy model to learn simple tasks in a Reinforcement Learning (RL) setting show that the model fails here as well. Experiments on this robot in Higuera et al. (2018), using the same reward formulation but a different learning mechanism, have shown that these tasks can be learned. Thus we attribute the breakdown of diversity learning to the inadequacy of the policy model.

2. Motivation

Recent work by Eysenbach et al. (2018) has demonstrated that by simultaneously learning a family of policies with the objective of maximizing their distinctiveness from one another in the state space, the resulting policies, called "skills" in that work, will guide an agent into diverse regions of the state space. These skills can then provide advantageous initial policies for reinforcement learning tasks. Their method, called Diversity Is All You Need (DIAYN), was demonstrated with simple legged agents on walking and running tasks in a simple physical simulation.

DIAYN trains a discriminator in parallel with its skills to determine their distinctiveness and supply the reward signal. When DIAYN is progressing as intended, the value of the reward achieved by the skills increases as their diversity, in terms of the state spaces they achieve, increases. Our intuition was that this reward, which functions as a metric of state space diversity, could prove to be a useful metric for informing the hardware design of a robotic agent.



a) Neutral Pose Configuration b) Knife-Edge Pose Configuration

Figure 1: Aqua in the Gazebo simulator.

When designing a robot, two qualities that often concern us are **flexibility**: the robot's ability to perform diverse tasks as required, and **robustness**: the robot's ability to persist or recover in the event of some catastrophe, such as hardware failure. One way of increasing the robustness of a robot is to add redundancy to its design by adding more end-effectors (such as arms or legs). This may benefit flexibility as well, but there are diminishing returns here: two arms may be much more flexible than one, but eleven arms may not be much more flexible than ten.

Having a quantitative metric of flexibility would allow us to apply this reasoning rigorously, and measure precisely the change in flexibility resulting from any change to a robot's design. A large increase in the metric for a given change implies that it adds significant flexibility; a small increase implies it primarily adds robustness. One way to obtain such a metric may be to run DIAYN training on different robot configurations, and observe the discriminator reward to which the training finally converges.

3. Method

3.1. Gazebo Underwater Aqua Simulator

Given the unforgiving and time-consuming nature of real underwater experiments, we conduct our experiments on a simulated Aqua robot. Our simulated Aqua is running in an extension of the simulation environment described in Meger et al. (2015) on top of the Gazebo simulator, first presented in Koenig and Howard (2004). Linear and rotational drag, mass, buoyancy and gravity are all accounted for, including drag and lift effects on each individual fin.

The state space used in our learning experiments has 7 dimensions consisting of the robot's orientation represented as roll (θ), pitch (ϕ), yaw (ψ), its depth, and its angular velocities ($\dot{\theta}$, $\dot{\phi}$, $\dot{\psi}$).

3.2. DIAYN

In the learning system used in Eysenbach et al. (2018), each skill consists of a policy network, a Q-function network, and Value-function network. These are all trained with respect to a reward produced by a discriminator network, which is trained in parallel with the skills. We make use of the same learning system in this work, with the difference that we use networks

consisting of 4 layers of 200 neurons each, vs 2 layers of 300 neurons in the original. We make this change to attempt to account for the more complex action space and dynamics of our environment. We used the algorithmic framework provided by Duan et al. (2016) to implement this section.

3.3. Fin Ablation Experiments

The Aqua robot is equipped with six individually-driven fins. The variable of robot design that we consider in our experiments is the number of fins available for control by the policy. For each configuration described below, we ran three DIAYN experiments, with the number of skills set to 10, 20, and 50.

In our various configurations, we treat each fin as completely independent from one another, with no coupling of any sort. Our configurations are the following:

- **Two-fin:** back two fins are actuated individually with action space of size [1x6].
- **Four-fin:** back two and front two fins are actuated individually with action space of size [1x12].
- **Six-fin:** all six fins are actuated individually with action space of size [1x18]

We run DIAYN on each configuration for three numbers of skills (10, 20, 50).

4. Results

4.1. Fin Ablation Results

The results of our fin-ablation experiments are presented in Figure 2. In all experiments we see an initially high reward from the still-uncertain discriminator; the reward quickly drops as the discriminator network learns to tell the difference between different states; and then, all going well, the reward should gradually recover as the skills grow more diverse.

As mentioned before, we had expected to observe increasing diversity reward for each additional pair of fins, but with a smaller jump from 4 to 6 than from 2 to 4. To our initial surprise, we observed the opposite: it is only with 2 fins that any significant diversity reward is achieved in any experiment, and 6 fins performs consistently worse than 4 fins. Also notable is that all three configurations fail to acheive much reward when 50 skills are learned simultaneously. This latter effect may simply be due to the space of skills being more crowded - with the same space being divided among more skills, we would expect discriminability between any two skills to be reduced on average.

4.2. Evaluation of the Model in an RL Setting

We suspected that the collapse of DIAYN learning with many fins may have been due to the complexity of the action space and the fact that this environment is a very challenging one, in comparison to those used in Eysenbach et al. (2018). Aqua's action space, as described in Section 3.1, is 6-D with two fins, 12-D with four, and 18-D with all six fins. Combined with the fact that these actions interact with each other in complex ways due to the fluid dynamics at play, and (perhaps most critically) the fact that some poorly-chosen initial

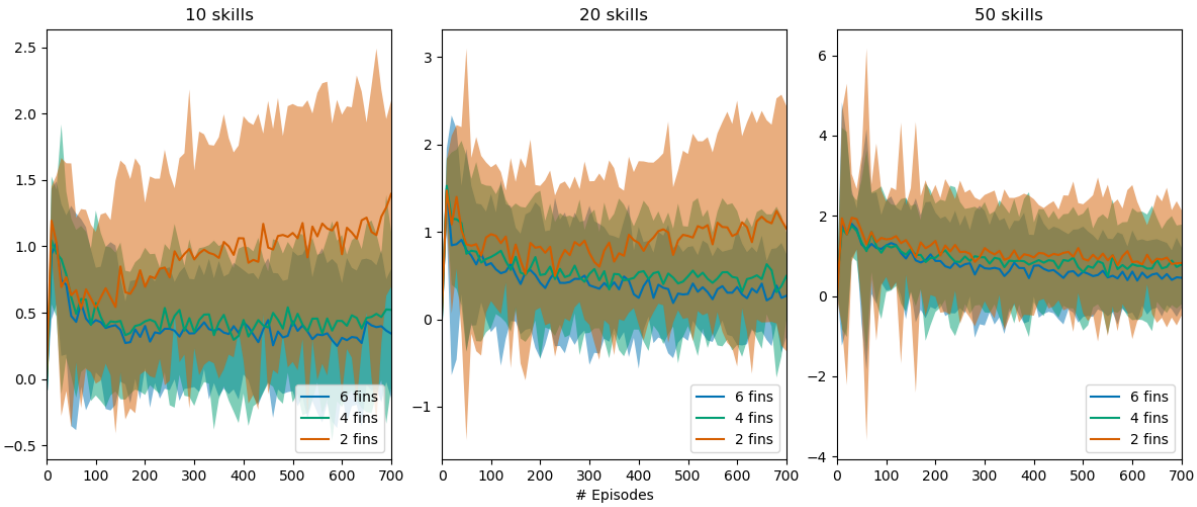


Figure 2: The per-episode diversity reward of the set of skills being trained at different numbers of fins, plotted at every 10th episode. We run DIAYN experiments with 10, 20, and 50 different skills being learned in parallel. At 10 and 20 skills, the 2-fin experiment manages to achieve some increase in reward; but the 4- and 6-fin experiments never do, and no fin configuration manages to at 50 skills.

actions can lead the robot into a "tumbling" motion that is difficult to escape from, this makes for a very difficult task. The model being used to learn the policy may simply not have been sophisticated enough to handle this complexity.

To test this conjecture, we applied the same learning model to the knife-edge swimming task, a simple RL task in which the robot is rewarded for rotating 90° about its roll axis and swimming forward while maintaining this orientation, as shown in Fig. 1. We again ran experiments with 2, 4, and 6 fins, and in each experiment we began with randomly-initialized V, Q, and policy networks, and trained the system for 700 episodes. Fig. 3 shows the results of these experiments. The agent fails to achieve any consistent cumulative reward over the training period, even with just two fins active. We take this as evidence that our hypothesis on the failure of DIAYN learning is correct: the learning architecture used here is not appropriate to the task at hand.

5. Conclusions and Future Work

In this work, we run the unsupervised skill-learning technique DIAYN on a realistic robot in a realistic and highly complex simulated environment. We assessed whether the central reward mechanism of DIAYN, the discriminator reward, would be useful as a comparative metric of the flexibility and robustness of different robot designs. Our results on this front are inconclusive, however. The model-free learning system we used as the backbone for this process, ironically, was itself not flexible enough to learn to control the robot in this

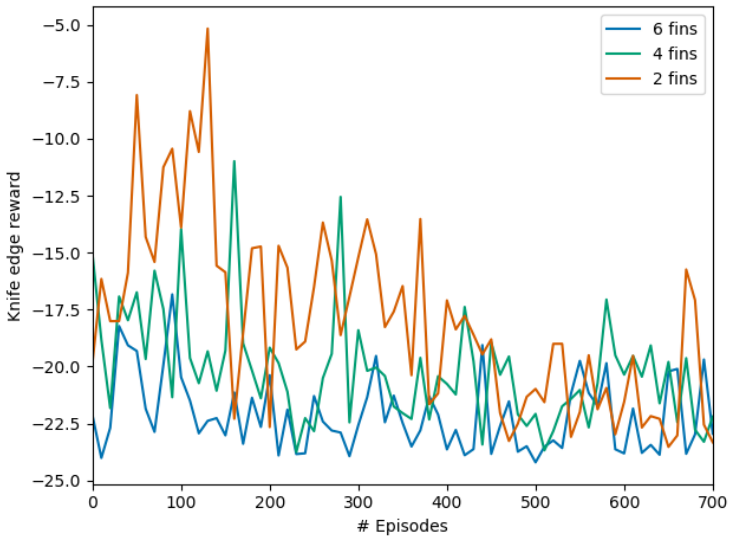


Figure 3: The cumulative episodic rewards recieved over 700 episodes of training, evaluated every 10 episodes, for all three configurations of fins. No improvement is observed in any case from the random initial policy,

challenging environment. To answer our research question, we will first have to run DIAYN with a more capable underlying learning system.

As detailed in Higuera et al. (2018), a model-based learning system has been shown to be very successful in learning simple tasks like knife-edge on the Aqua robot, both in simulation and in the real world. The application of this model-based learning system, in place of the model-free architecture used in these experiments, will give the DIAYN technique a much stronger foundation in this environment. We anticipate pursuing model-based learning with DIAYN on Aqua as a direction of future work.

References

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *CoRR*, abs/1802.06070, 2018. URL http://arxiv.org/abs/1802.06070.

Juan Camilo Gamboa Higuera, David Meger, and Gregory Dudek. Synthesizing neural network controllers with probabilistic model based reinforcement learning, 2018.

N. Koenig and A. Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and*

Systems (IROS) (IEEE Cat. No.04CH37566), volume 3, pages 2149–2154 vol.3, Sept 2004. doi: 10.1109/IROS.2004.1389727.

David Meger, Juan Camilo Gamboa Higuera, Anqi Xu, Philippe Giguere, and Gregory Dudek. Learning legged swimming gaits from experience. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 2332–2338. IEEE, 2015.