Investigating Deep Reinforcement Learning for Grasping Objects with an Anthropomorphic Hand

Mayur Mudigonda¹, Pulkit Agrawal², Michael R Deweese¹, Jitendra Malik²†

ABSTRACT

Grasping objects with high dimensional controllers such as an anthropomorphic hand using reinforcement learning is a challenging problem. In this work we experiment with a 16-D simulated version of a prosthetic hand developed for SouthHampton Hand Assessment Procedure (SHAP). We demonstrate that it is possible to learn successful grasp policies for an anthropomorphic hand from scratch using deep reinforcement learning. We find that our grasping model is robust to sensor noise, variations in object shape, position of the object and physical parameters such as the density of the object. Under these variations, we also investigate the utility of touch sensing for grasping objects. We believe that our results and analysis provide useful insights and strong baselines for future research into the exciting direction of object manipulation with anthropomorphic hands using proprioceptive and other sensory feedback.

1 INTRODUCTION

Grasping is a basic building block for object manipulation tasks. It is a well established and important problem in robotics that has garnered a lot of research interest (11, 12, 13, 14, 15). While past works have made excellent progress in demonstrating the ability to grasp from high-dimensional sensory observations such as images, they have employed rather simplistic two finger parallel jaw grippers. Pragmatically, it can be argued that if the end goal is simply to grasp an object of interest, parallel jaw grippers are sufficient for grasping most objects. However just grasping an object is not a very useful skill by itself and humans often perform complex in-hand manipulation of grasped objects for achieving desired goals. The ability to perform human like in-hand manipulation has generated a lot of interest in design and control of anthropomorphic hands such as the Shadow Hand, the ADROIT suite (12), the UW Hand (13), RBO 2 hand (14) and many others (15, 16).

The superiority of anthropomorphic hands to parallel jaw grippers for in-hand manipulation comes at the cost of difficulty in finding a policy to control them in a desirable fashion. Past work on controlling anthropomorphic hands has relied on imitation (15), human demonstration (17, 18) or pre-defined grasp types (15) as priors for inferring control strategies. Another line of work has exploited compliance and under-actuation (19, 20) for reducing the number of degrees of actuation which in turn simplifies the control problem. However compliance inevitably leads to loss in precise knowledge of the object state. As the observation space becomes richer and the complexity of tasks increases, constructing analytical sensorimotor models becomes progressively harder and using human demonstrations for learning control policies is not scalable because collecting demonstrations is known to be a hard and tedious process.

Reinforcement learning provides a general paradigm for determining control policies without explicitly modeling the actuator or the environment. However, most current reinforcement learning algorithms (21, 22, 23, 24) rely on random walk in the action space for exploration which makes it non-trivial to control actuators with large number of degrees of freedom. It is therefore not surprising and to the best of our knowledge, learning robust grasping behavior from scratch using model free reinforcement learning with a joint controlled anthropomorphic hand in the face of changing environmental conditions and sensory noise has not yet been demonstrated. Given that it is quite challenging to manipulate objects with an anthropomorphic hand, in this work we start with learning the grasping primitive and investigate the limitations and successes of current reinforcement learning techniques towards this end.

2 EXPERIMENTAL SETUP

Environment and Observations: We use a simulated model of the anthropomorphic hand used as part of SouthHampton Hand Assessment Procedure (SHAP). We demonstrate that it is possible to learn successful grasp policies for an anthropomorphic hand from scratch using deep reinforcement learning. We find that our grasping model is robust to sensor noise, variations in object shape, position of the object and physical parameters such as the density of the object. Under these variations, we also investigate the utility of touch sensing for grasping objects. We believe that our results and analysis provide useful insights and strong baselines for future research into the exciting direction of object manipulation with anthropomorphic hands using proprioceptive and other sensory feedback.
observation space (input to the model) consisted of the internal position and velocity of the 25 joints and the object position resulting in 53 dimensions (25*2 + 3 = 53). In some experiments, we made use of the touch sensors (19 in number); in these cases, the total number of dimensions were 72. In all experiments, this hand was tasked to grasp a single object kept on a table.

**Reward Structure:** order to grasp an object the hand must reach the object and then lift it. One way to specify the reward for grasping is the height of the object from the table ($D_{ot}$). However, this reward is sparse because random exploration is unlikely to result in object grasps by chance. Unsurprisingly, with this sparse reward, learning was unsuccessful. In order to encourage the hand to grasp the object we shaped the reward to penalize the distance from the palm of the hand to the object (i.e. $-D_{op}$). Our reward function therefore looks like: $r(t) = -0.1 \times D_{op}$ where $D_{op} \geq \epsilon$ for reaching and $r(t) = D_{ot}$ where $D_{op} < \epsilon$ for grasping. Where $\epsilon$ is the parameter that encourages the hand to reach the object. We set $\epsilon$ to be 0.15, which was about approximately two times the radius of the sphere object used in our experiments. A positive value signifies that the hand has reached the object and is manipulating it. A value of approximately 120 and above implies the object was grasped and lifted off of the table successfully.

### 3 Results

We used trust region policy optimization (TRPO) (22) as the policy learning algorithm. We found that it was hard to train the off-policy methods such of deep deterministic policy gradient (DDPG) (21) for this task. For TRPO, the neural network policy was represented by a multi-layer perceptron (MLP) with two hidden layers with sizes of 32 units and 32 units which serve to model a Gaussian distribution from which we sample actions. We experimented with varying the size of the network by testing networks of size 64 and 16 but found that they performed comparably (64 units) or worse (16 units). Based on these preliminary results we used two layer neural networks of size (32, 32) for the rest of our experiments.

We found that three parameters were critical to training: (a) batch size (b) the initial standard deviation of the Gaussian policy and (c) normalizing the observations. We varied the initial standard deviation of the Gaussian across multiple experiments and settled on a value of 3.0. We fixed the episode length to be 500 time steps long. We empirically found that this was a long enough trajectory to reach to almost any part of the table with the actions that were being generated by the policy. To normalize the observations, we unrolled 1000 random episodes and computed their means and standard deviations. These were then used to normalize the observations during learning. We tried experiments without normalizing the observations and the model failed to learn. Given the seed sensitivity of many Deep-RL algorithms, we were cautious and ran multiple (at least three seeds) for each experiment.

#### 3.1 Batch Size

To test the robustness of training to batch size, we used batch sizes of - 40k, 60k, 80k and 120k. As batch size was increased the model learnt a more robust policy for grasping and lifting the object. The performance scaled linearly with the number of trajectories TRPO saw. The numbers in Table 1 describe the cumulative reward (undiscounted) averaged across multiple trajectories.

<table>
<thead>
<tr>
<th>batch size</th>
<th>feature space</th>
</tr>
</thead>
<tbody>
<tr>
<td>40k</td>
<td>169.499 +/- 96.98</td>
</tr>
<tr>
<td>60k</td>
<td>332.57 +/- 21.04</td>
</tr>
<tr>
<td>80k</td>
<td>281.887 +/- 36.65</td>
</tr>
<tr>
<td>120k</td>
<td>349.90 +/- 38.60</td>
</tr>
</tbody>
</table>

Table 1: Batch size

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Sphere Success Rate</th>
<th>Multi-Obj Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuboid</td>
<td>0.92</td>
<td>1.0</td>
</tr>
<tr>
<td>sphere</td>
<td>0.85</td>
<td>0.99</td>
</tr>
<tr>
<td>ellipsoid</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>cylinder</td>
<td>0.17</td>
<td>1.0</td>
</tr>
<tr>
<td>can</td>
<td>0.0</td>
<td>0.06</td>
</tr>
<tr>
<td>coin</td>
<td>0.46</td>
<td>0.47</td>
</tr>
<tr>
<td>screwdriver</td>
<td>0.29</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 2: Multiple Objects - Generalization

Table 1 comparing the mean and standard deviation (across three seeds) average reward per episode during training for different batch sizes. Table 2 describing the success of grasping an object and lifting it off the table. Success rate is defined here as the number of times the object was at a distance of 0.7 or greater across all episodes for that object. In our simulation this usually meant the object must have been grasped reasonably for the object to be that far from the table. We simulated 1000 episodes and this led approximately to 125 episodes per category.

#### 3.2 Multiple Objects and Generalization

Once we found that it was possible to grasp a single object (sphere), we sought to investigate if it was possible to grasp previously unseen objects. This resulted in poor generalization as seen in column 1 of Table 2. Notably, although the model was trained on spheres it generalized better to cuboids due to their relative simplicity in grasping but did poorly on all other objects.

Given the poor generalization with a single object during training, we then setup an experiment with multiple objects - a sphere, cuboid and an ellipsoid. We gave the learnt model novel objects such as a screw driver, a can, an ellipsoid and a cylinder. We show that we are able to learn a policy to grasp for multiple objects as well. Some visualizations from this experiment can be seen in Figure 1. We see that it learns to grasp multiple objects it had not seen during training. Nevertheless, it was not always successful; often failing with objects that did not share any common geometric properties with objects in training.
3.3 Generalization to Variation in Object Density and Sensory Noise

The mainstream mechanism for obtaining generalization is to train a policy on different instances of the same problem, where one or more parameters (such as density, sensory noise) of interest are varied. The hope is that finding a policy that works across variations in training set will also work for test set. While, this might be reasonable if the test set merely requires interpolation between two points in the training set, it is unlikely to be adequate for extrapolative generalization. We tested this in two setting of varying the density of the object and the amount of sensory noise.

In the first set of experiments, we trained one model with mass variation from 1 to 2 (blue) and other model with variation from 1 to 3 (orange). We tested these models on objects whose densities were sampled between 2 and 3 (see Figure 2a). If simply varying the density is sufficient for learning a robust policy, the performance gap between the two trained models should be small. Figure 2a shows that this is not the case. Further analysis reported in Figure 2b suggests that the average force exerted by the hand while lifting the object is higher when the model is trained with larger density variations (orange) as compared to training with smaller density variations (blue), suggesting that the trained policy can adapt to heavier objects to some extent. However, as witnessed from Figure 2a the extent of this adaptation is insufficient.

In a second set of experiments, we trained these models with a noise standard deviation $\sigma = 1$. During test time, we sampled noise from a uniform distribution $\eta \in (0, 2)$. Figure 2c suggests that the the performance deteriorates gracefully with more noise. We next sought to investigate if including touch sensing would increase the robustness of grasping. We re-ran this experiment with the observation space expanded to include the readings of 19 pressure sensors on the SHAP MPL hand used in our experiments. Figure 2c shows that the policy using touch sensing (orange) performs significantly better across all noise levels.

4 Conclusion

In this work we show that it is possible to learn robust grasping policies from anthropomorphic hands using model free reinforcement learning algorithm known as trust region policy optimization (TRPO). We found that the learned policies were not object specific and when trained with three objects, our system was able to grasp four novel objects with quite different geometries as compared to the training set of objects. Further investigation revealed that the learned policies were robust to some extent to noise in sensory observations and variation in object properties such as density. The results pointed that simple randomization of parameters is insufficient for extrapolative generalization. Finally, we found that policies trained with touch sensing were more robust to sensory noise than policies trained without touch sensing.
REFERENCES


