High Dimensional Representations for Learning Grasping in Clutter Policies from Demonstrations

Michael Laskey*, Chris Power*, Ken Goldberg*
*Dept. of EECS
University of California, Berkeley
Email: { laskeymd, chris.powers, goldberg } @berkeley.edu

I. Introduction

Recent advances in deep learning have enabled multiple successes in learning manipulation policies from high dimensional image data [2, 3]. Despite these successes, high dimensional Learning from Demonstrations (LfD) raises an interesting statistical challenge. In general, most LfD applications have small or limited data sets due to the cost of collection. However, high-dimensional state representation can require training deep policies with potentially more parameters than the number of available examples, which is concerning for generalization.

One approach to alleviate this challenge is to choose a state representation that is sufficient to perform the task, but is not unneccsarily large. By reducing the state space size, we can reduce the parameters in the model and improve generalization [8]. However, it is not clear in general, what information is important for learning manipulation policies. For example, if a robot wants to push objects should it consider only object shape or also object texture? In this paper, we explore various representations for the task of grasping in clutter.

We specifically consider a form of grasping in clutter inspired by the Amazon Picking Challenge [1]. A robot is shown a set of of objects on a shelf and must clear a path to reach a goal object. The robot's perspective is from a moving eye-in-hand viewpoint, shown in Fig. 1, which forces the robot to learn to search for the object in high-dimensional image space. In our setup, we experimentally find that reduce representations, such as a binary image representation (shown in Fig 2B), can lead to better generalization than standard RGB-D representations taken from this viewpoint.

II. Grasping in Clutter System

We used a Toyota HSR robot to perform the grasping in clutter task. The objects, shown in Fig. 1, are common household food items. The target object, a mustard bottle, is placed behind the other objects and is occluded from the robot's viewpoint. Thus requiring the robot to search for the goal object.

We collect demonstrations via tele-operation with an Xbox controller. The robot's motion is constrained to its mobile base, which has three degrees of freedom translation and rotation, or$x$, y and θ. During tele-operation the supervisor sends change in position commands to these three degrees of freedom. During a demonstration, RGB-D data is recorded from a primesense structured light sensor, which is mounted at its head. The images are cropped to obtain the viewpoint shown in Fig. 2, which is similar to an eye-in-hand view.

Fig. 1: Experimental setup for grasping in clutter. A Toyata HSR robot uses its arm to push obstacle objects out of the way to reach the goal object, which is a mustard bottle.The robot's policy for pushing object's away is represented as a neural network trained on images taken from the robot's primesense camera. The cropped viewpoint is of the image is shown in the orange box.

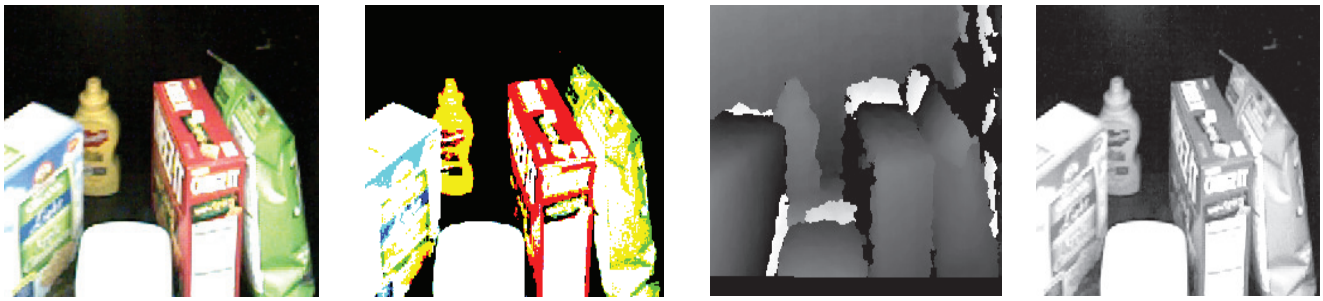
During data collection the four obstacle objects are rotated in pose and relative position. We collected 60 demonstrations from a human supervisor, where each demonstration ends once the mustard bottle is clearly visible from the robot's camera. In total it took 1 hour and 10 minutes to collect the training data.

Our learned policy class is a neural network architecture consisting of a convolutional layer, a pooling layer and two fully connected. The architecture is inspired by that used in [2]. During training, we regress from the chosen state space to the control vector, which is a change in position. Our loss function is a squared euclidean loss. We trained the network using Tensorflow on a Tesla K-40 GPU with a Momentum optimization scheme.

III. State Representations

We consider two techniques to reduce data needed for learning: 1) changing the state space representation to reduce the dimensionality and 2) providing synthetic data augmentation to increase robustness to lighting changes.

Representations We consider six different representations for the grasping in clutter task. The first representation is RGB images, or full color, which has been shown to be sucessful for manipulation tasks [4]. Another representation used for grasping in clutter is a binary mask over the RGB image [2].



A) RGB B) Binary C) Depth D) Gray

Fig. 2: Different state representations applied to the images from the robot's primesense camera: A) RGB color images. B) RGB images with a binary mask C) Depth images and D) Grayscale images.

The proposed binary mask augments each channel of the image by setting each pixel values to 255 or 0 based on whether that pixel is above or below a specified threshold. We chose the threshold of 180 based on manual tuning for high contrast. An example image of the binary mask can be seen in Fig. 2B. We also consider grayscale images, which also reduce the effective state space of the RGB image.

In addition to color data, it has been common in the literature to also include depth information [3], which can be obtained from a structured light sensor. We consider both RGB-D states spaces which are a 4 channel image and also a depth only representation, which has been shown to be sufficient for grasping [5]

Finally, instead of learning convolutional filters, we also consider pre-trained convolutional layers from the VGG architecture trained for image classification on ImageNet [7]. Our intuition is that natural images have common statistics and features learned over a very diverse set may be sufficient for the grasping in clutter task.

Data Augmentation Another technique that has been shown to increase the robustness of the learned features is to add artificial lighting changes to the training images. The intuition is that by forcing the network to be robust to lighting changes it may lead to a representation that have higher invariance to small pixel perturbations.

We examine applying variations in lighting via modifying intensity. During training of the network we apply these variations to the training data, which increased the dataset by 4x.

IV. Experiments

We perform two experiments to test the generalization of the state representations. The first is examining the error on a held out test set and the second is validation on the physical robot of the best state representation.

Representations In order to test generalization of the different representations, we divided our 60 demonstration dataset into 48 demonstrations for the training set and 12 demonstrations for test set, which corresponds to 20%. We then varied the state space representation and trained each neural network for a fixed 3000 iterations. We report the lowest squared euclidean loss achieved during training in Table 1.

Our results suggest that the binary masks lead to better generalization with 0.05 loss on test. However, VGG features are able to get much lower training error than the other state representations. We found that data augmentation did not help generalization. In general, the training error is higher with data augmentation than without and subsequently the test error is higher. A more expressive network may be needed to learn the lighting invariance.

	No D.A.		D.A.	
	Train	Test	Train	Test
VGG	**0.002**	0.06	0.05	0.06
RGB-D	0.03	0.06	0.04	0.07
Color	0.04	0.06	0.11	0.05
Gray	0.10	0.10	0.11	0.12
Depth	0.08	0.08	N.A.	N.A.
Binary	0.03	**0.05**	0.07	0.07

TABLE I: Squared Euclidean error for different state representations on a dataset of 60 demonstrations of grasping in clutter. VGG features with no Data Augmentation (D.A.) achieve the lowest training error, while Binary masks receive the lowest test error. Thus, suggesting binary masks can help increase generalization.

Physical Robot We next evaluated the network trained with the binary masks on the physical robot. To determine success of identifying the mustard, we use the Faster R-CNN, which is trained to detect bottles on the Pascal VOC vision dataset [6]. During policy roll-out, we query the R-CNN at each timestep to determine if the policy is successful. Once the mustard is detected the robot leverages a motion planner to reach the target object.

We evaluated the policy on 15 test configurations, which are drawn from a similar initial state distribution as training. The policy was successful 86% of time in retrieving the mustard. The common failure modes appear to be due to the covariate shift error, where the robot drifts from the demonstrations and is unable to recover. In future work, we hope to explore how injecting noise in the demonstrations can alleviate this problem.

References

[1] Nikolaus Correll, Kostas E Bekris, Dmitry Berenson, Oliver Brock, Albert Causo, Kris Hauser, Kei Okada, Alberto Rodriguez, Joseph M Romano, and Peter R Wurman. Lessons from the amazon picking challenge. *arXiv preprint arXiv:1601.05484*, 2016.

[2] Michael Laskey, Jonathan Lee, Caleb Chuck, David Gealy, Wesley Hsieh, Florian T Pokorny, Anca D Dragan, and Ken Goldberg. Robot grasping in clutter: Using a hierarchy of supervisors for learning from demonstrations.

[3] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *arXiv preprint arXiv:1504.00702*, 2015.

[4] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *arXiv preprint arXiv:1603.02199*, 2016.

[5] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] V Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.