End-to-End Training of Deep Visuomotor Policies

Sergey Levine*, Chelsea Finn*, Trevor Darrell, and Pieter Abbeel
Department of Electrical Engineering and Computer Science
University of California, Berkeley, Berkeley, CA 94709
{svlevine, cbfinn, darrell, pabbeel}@cs.berkeley.edu

I. Overview

Policy search methods can allow robots to autonomously learn a wide variety of behaviors. However, policies learned using such methods often rely on hand-engineered components for perception and low-level control. For example, a policy for object manipulation might specify motions in task-space, using hand-designed PD controllers to execute the desired motion and relying on an existing vision system to localize objects in the scene [5]. The vision system in particular can be complex and prone to errors, and its performance is typically not improved during policy training, nor adapted to the goal of the task.

We propose a method for learning policies that directly map camera images and joint angles to motor torques. The policies are trained end-to-end using real-world experience, optimizing both the control and vision components on the same measure of task performance. This allows the policy to learn goal-driven perception, which avoids the mistakes that are most costly for task performance. Learning perception and control in a general and flexible way requires a large, expressive model. We use convolutional neural networks (CNNs), which have 92,000 parameters and 7 layers. Deep CNN models have achieved state of the art results on supervised vision tasks [1, 6], but sensorimotor deep learning remains a challenging prospect. The policies are extremely high dimensional, and the control task is partially observed, since part of the state must be inferred from images.

To address these challenges, we extend the framework of guided policy search to sensorimotor deep learning. Guided policy search decomposes policy search into two phases: a trajectory optimization phase that determines how to solve the task in a few specific conditions, and a supervised learning phase that trains the policy from these successful executions with supervised learning [3]. Since the CNN policy is trained with supervised learning, we can use the tools developed in the deep learning community to make this phase simple and efficient. We handle the partial observability of visuomotor control by optimizing the trajectories with full state information, while providing only partial observations (consisting of images and robot configurations) to the policy. The trajectories are optimized under unknown dynamics, using real-world experience and minimal prior knowledge.

The main contribution of our work is a method for end-to-end training of deep visuomotor policies for robotic manipulation. This includes a partially observed guided policy search algorithm that can train high-dimensional policies for tasks where part of the state must be determined from camera images, as well as a novel CNN architecture designed for robotic control, shown in Figure 1. Our results demonstrate improvements in consistency and generalization from training visuomotor policies end-to-end, when compared to the more standard approach of training the vision and control components separately. A complete description of our work can be found in our recent technical report [2], and videos of the learned policies can be found on the project website: [http://sites.google.com/site/visuomotorpolicy].

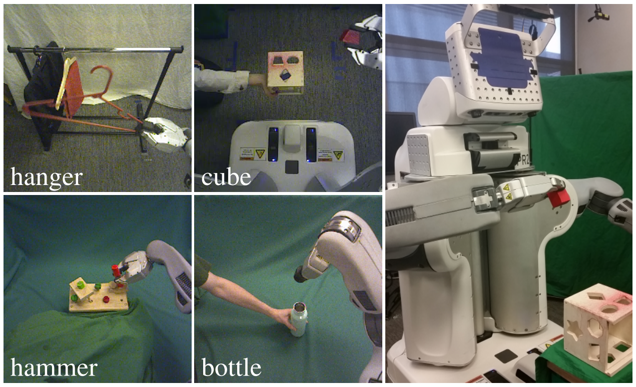

Fig. 1: Our method learns visuomotor policies that directly use camera image observations (left) to set motor torques on a PR2 robot (right).

II. Experimental Results

We evaluated our method by training policies on a PR2 robot for hanging a coat hanger on a clothes rack, inserting a block into a shape sorting cube, fitting the claw of a toy hammer under a nail with various grasps, and screwing on a bottle cap (see Figure 1). The cost function for these tasks encourages low distance between three points on the end-effector and corresponding target points, low torques, and, for the bottle task, spinning the wrist. The equations for these cost functions follow prior work [3]. Each task involved variation of about 10-20 cm in the horizontal position of the target object (the rack, shape sorting cube, nail, and bottle). The coat hanger and hammer tasks were also trained with two and three grasps, respectively. All tasks used the same policy architecture.

We evaluated the visuomotor policies in three conditions: (1) the training target positions and grasps, (2) new target positions not seen during training and, for the hammer, new

Asterisk (*) denotes authors who contributed equally to the work.

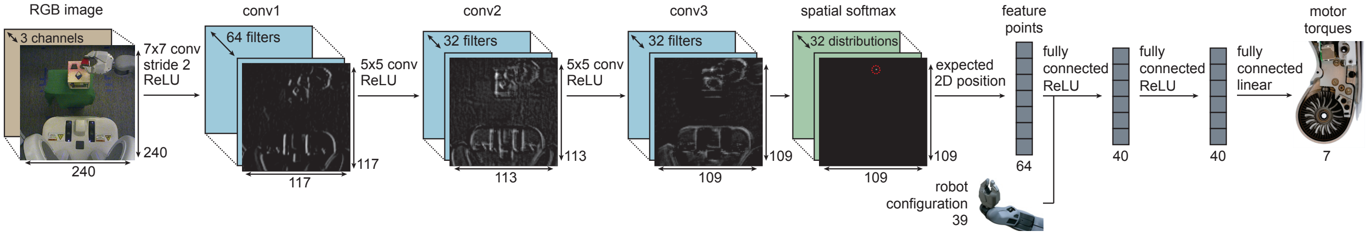

Fig. 2: Visuomotor policy architecture. The network contains three convolutional layers, followed by a spatial softmax and an expected position layer that converts pixel-wise features to feature points, which are better suited for spatial computations. The points are concatenated with the robot configuration, then passed through three fully connected layers to produce the torques.

grasps (spatial test), and (3) training positions with visual distractors (visual test). A selection of these experiments can be viewed in the supplementary videos. For the visual test, the shape sorting cube rested on a table, the coat hanger was placed on a rack with clothes, and the bottle and hammer tasks were performed in the presence of clutter.

The success rates for each test are shown in Table I. We compared to two baselines, both of which train the vision layers in advance for pose prediction, instead of training the entire policy end-to-end. The features baseline discards the last layer of the pose predictor and uses the feature points, resulting in the same architecture as our policy, while the prediction baseline feeds the predicted pose into the control layers.

The pose prediction baseline is analogous to a standard modular approach to policy learning, where the vision system is first trained to localize the target, and the policy is trained on top of it. This variant achieves poor performance, because although the pose is accurate to about 1 cm, this is insufficient for such precise tasks. As shown in the supplementary video, the shape sorting cube and bottle cap insertions have tolerances of just a few millimeters. Such accuracy is difficult to achieve even with calibrated cameras and checkerboards. Indeed, prior work has reported that the PR2 can maintain a camera to end effector accuracy of about 2 cm during open loop motion [4]. This suggests that the failure of this baseline is not atypical, and that our visuomotor policies are learning visual features and control strategies that improve the robot's accuracy.

When provided with pose estimation features, the policy has more freedom in how it uses the visual information, and achieves somewhat higher success rates. However, full end-to-end training performs significantly better, achieving high accuracy even on the challenging bottle task, and successfully adapting to the variety of grasps in the hammer task. This suggests that, although the vision layer pre-training is clearly beneficial for reducing computation time, it is not sufficient by itself for discovering good features for visuomotor policies.

The policies exhibit moderate tolerance to distractors that are visually separated from the target object. However, as expected, they tend to perform poorly under drastic changes to the backdrop, or when the distractors are adjacent to or occluding the manipulated objects, as shown in the supplementary videos. In future work, this could be mitigated by varying the scene at training time, or by artificially augmenting the image samples in the training set with synthetic transformations.

coat hanger	training (18)	spatial test (24)	visual test (18)
end-to-end training	**100%**	**100%**	**100%**
pose features	88.9%	87.5%	83.3%
pose prediction	55.6%	58.3%	66.7%
shape sorting cube	training (27)	spatial test (36)	visual test (40)
end-to-end training	**96.3%**	**91.7%**	**87.5%**
pose features	70.4%	83.3%	40%
pose prediction	0%	0%	n/a
toy claw hammer	training (45)	spatial test (60)	visual test (60)
end-to-end training	**91.1%**	**86.7%**	**78.3%**
pose features	62.2%	75.0%	53.3%
pose prediction	8.9%	18.3%	n/a
bottle cap	training (27)	spatial test (12)	visual test (40)
end-to-end training	**88.9%**	**83.3%**	**62.5%**
pose features	55.6%	58.3%	27.5%

TABLE I: Success rates on training positions, on novel test positions, and in the presence of visual distractors. The number of trials per test is shown in parentheses.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. 2012.

[2] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *arXiv preprint arXiv:1504.00702*, 2015.

[3] S. Levine, N. Wagener, and P. Abbeel. Learning contact-rich manipulation skills with guided policy search. In *International Conference on Robotics and Automation (ICRA)*, 2015.

[4] W. Meeussen, M. Wise, S. Glaser, S. Chitta, C. McGann, P. Mihelich, E. Marder-Eppstein, M. Muja, Victor Eruhimov, T. Foote, J. Hsu, R.B. Rusu, B. Marthi, G. Bradski, K. Konolige, B. Gerkey, and E. Berger. Autonomous door opening and plugging in with a personal robot. In *International Conference on Robotics and Automation (ICRA)*, 2010.

[5] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal. Learning and generalization of motor skills by learning from demonstration. In *International Conference on Robotics and Automation (ICRA)*, 2009.

[6] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.