A Systematic Practice of Judging the Success of a Robotic Grasp Using Convolutional Neural Network

Hengshuang Liu, Pengcheng Ai, Junling Chen
College of Physical Science and Technology, Central China Normal University
NO.152 Luoyu Road, Wuhan, Hubei, P.R.China 430079
{hengshuangliu, pengcheng.ai, cjl1213}@mails.ccnu.edu.cn

Abstract

In this abstract, we present a novel method using the deep convolutional neural network combined with traditional mechanical control techniques to solve the problem of determining whether a robotic grasp is successful or not. To finish the task, we construct a data acquisition platform capable of robot arm grasping and photo capturing, and collect a diversity of pictures by adjusting the shape and posture of the objects and controlling the robot arm to move randomly. For the purpose of validating the generalization capability, we adopt a stochastic sampling method based on cross validation to test our model. The experiment shows that, with an increasing number of shapes of objects involved in training, the network can identify new samples in a more accurate and steadier way. The accuracy rises from 89.2% when we use only one category of shape for training to above 99.7% when we use 17 categories for training.

Introduction

In the domain of robotics and artificial intelligence, grasping an object by mechanical devices is an important issue for scientists. Sergey Levine et al. in Google presented a method in their paper about hand-eye coordination for robotic grasping, which used a combination of sensors and traditional image identification techniques (Levine et al. 2016) to properly function. In tradition, specialized mechanical structures and man-made criteria are needed for prompt sensing and accurate judgment (Farrow and Correll 2015; Konstantinova, Stilli, and Althoefer 2015). However, if we can simplify utilities and make the system reliable, it will contribute to the popularity of robot arms considerably.

In recent years, deep learning has achieved excellent results in image classification tasks (Krizhevsky, Sutskever, and Hinton 2012; Szegedy et al. 2015). Similarly, we consider using deep learning in the problem of judging the success of a robotic grasp. The model we use is the convolutional neural network, which has an input layer of a single picture taken from a camera. In the model, convolution layers and full-connected layers are built to extract features from the original picture. When training, labels are set and cross entropy is computed for back propagation. In the test, we compare the expected results from the model to the ground-truth results to get the statistics of the accuracy.

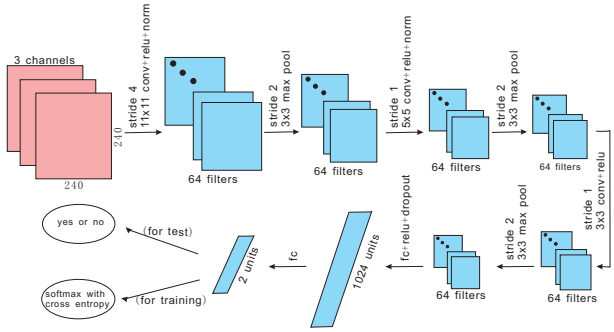

Figure 1: The structure of the convolutional neural network

Our approach collects pictures which have slightly different features by changing the shape and posture of the objects. Meanwhile, we control the robot arm to move randomly in a limited space, which results in pictures with changing backgrounds. Pictures are collected in a similar way when the robot arm has no load.

Modeling Method

Our model is motivated by commonly used neural network models, such as LeNet, AlexNet and GoogleNet. From Figure 1, we can see that the size of the input picture is 240×240, with 3 different color channels, and the hidden layers include 3 convolution layers and 2 full-connected layers. The core operation in the model is convolution, which is designed to utilize the property of translation invariance in computer vision.

There are some intricate factors involved in the implementation of the network. In the initialization process, the weights are initialized to Gaussian distribution with 0 mean and a small deviation, while the bias is initialized to 0 or a small value. To augment our dataset with limited samples (Gan et al. 2015), we randomly crop a 240×240 region from an original 240×320 picture and send it into our network for training. The dropout ratio is adjustable. We set it to 0.5, which is recommended to be used with a hidden layer. When implementing our network, all the convolution layers and Max Pool layers use a round down strategy implicitly

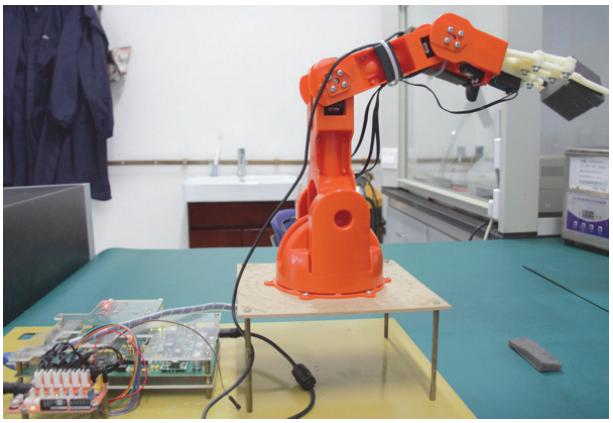

Figure 2: A side view of our data acquisition platform

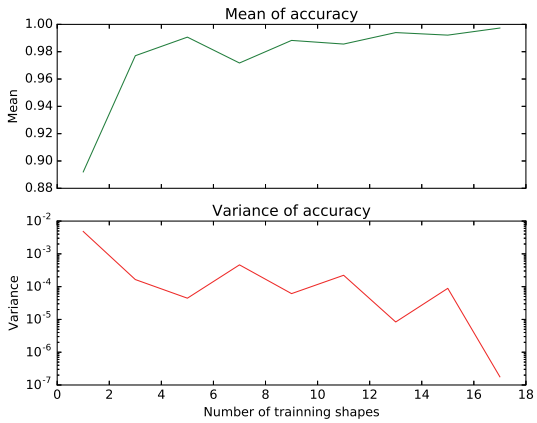

Figure 3: The mean and variance of accuracy (first test)

adopted by our software package.

Data Acquisition Platform

We construct a simple and effective data acquisition platform. The robot arm we use is the Arduino Braccio suite. It has 6 degrees of freedom, in which 5 degrees control 5 rotation axes to rotate from $0°$ to $180°$ and 1 degree controls the grasping action of the gripper. Every degree of freedom is driven by a servo. The circuit board used to control the system is the HPDAQ data acquisition board with a Xilinx Kintex 7 FPGA on it. HPDAQ outputs 6 channels of PWM wave, each of which directs a servo's movement. At the PC terminal, the host computer communicates with the HPDAQ via Ethernet using TCP/IP protocol. The HPDAQ decodes the control message sent by the computer into the high-level width of PWM wave in each channel. Our data acquisition platform is shown in Figure 2.

Experimental Details

We test our method in two ways. The first one is based on cross validation. As the structure of our model is fixed, the main purpose of the validation is to find out how the diversity of training data impacts on the generalization capability. We stochastically select some shapes of objects as our training set. The number of selected shapes ranges from 1 to 17 and each shape has 5 different postures. Next, we select another 3 shapes of objects from the remaining shapes as our validation set in a similar stochastic way. The non-grasping pictures are also divided into 20 groups, so we add the same number of non-grasping pictures into the training set and validation set. This procedure is repeated 5 times, and we can calculate the mean and variance of accuracy reflecting generalization capability. The result is shown in Figure 3.

The second way is using models resulted from different selected shapes in the first way to test on the 200 pictures (100 grasping pictures, 100 non-grasping pictures) in the test set. The grasping pictures in the test set have no identical pictures, differing either in shape or posture. The main characteristic of this way is that the test set is fixed for all models. The result is shown in Table 1.

Table 1: Experimental results (second test)

Number	Average accuracy	Variance of accuracy
1 shape	88.40%	6.84E-03
3 shapes	98.20%	1.58E-04
5 shapes	99.00%	2.50E-05
7 shapes	98.50%	2.88E-04
9 shapes	99.30%	7.50E-06
11 shapes	98.90%	1.80E-04
13 shapes	99.20%	7.50E-06
15 shapes	99.60%	5.00E-06
17 shapes	99.40%	5.00E-06

Conclusion

Deep learning approaches were heavily studied on several benchmarks to promote better theoretical works. However, its applications to issues combined with mechanics and robotics are just getting started. A point in the future is to apply our method in a multi-agent or diagnostic setting, where visual information should be sufficiently exploited.

References

Farrow, N., and Correll, N. 2015. A soft pneumatic actuator that can sense grasp and touch. In *IROS*, 2317–2323. IEEE.

Gan, Z.; Henao, R.; Carlson, D.; and Carin, L. 2015. Learning deep sigmoid belief networks with data augmentation. In *AISTATS*, 268–276.

Konstantinova, J.; Stilli, A.; and Althoefer, K. 2015. Force and proximity fingertip sensor to enhance grasping perception. In *IROS*, 2118–2123. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

Levine, S.; Pastor, P.; Krizhevsky, A.; and Quillen, D. 2016. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *arXiv preprint*.

Szegedy, C.; Liu, W.; Jia, Y.; et al. 2015. Going deeper with convolutions. In *CVPR*, 1–9.