Robotic Grasp Detection using Deep Convolutional Neural Networks

Authors: Sulabh Kumra and Christopher Kanan
Publisher: International Conference on Intelligent Robots and Systems (IROS)
Year: 2017

Keywords: grasp pose, DCNN, image-wise split, object-wise split, uni-modal, multi-modal

1. **Abstract:**

"Deep learning has significantly advanced computer vision and natural language processing. While there have been some successes in robotics using deep learning, it has not been widely adopted. In this paper, we present a novel robotic grasp detection system that predicts the best grasping pose of a parallel-plate robotic gripper for novel objects using the RGB-D image of the scene. The proposed model uses a deep convolutional neural network to extract features from the scene and then uses a shallow convolutional neural network to predict the grasp configuration for the object of interest. Our multi-modal model achieved an accuracy of 89.21% on the standard Cornell Grasp Dataset and runs at real-time speeds. This redefines the state-of-the-art for robotic grasp detection."

2. **Introduction:**
    - Humans have an excellent skills in grasping the objects in no time. To gain the same result from the robots is difficult.
    - It is stated that the state-of-the architecture for grasping would fail in real-time.
    - In general, the problem of grasping is classified into three types:
        - Grasp detection
        - Trajectory Planning
        - Execution
    - In the first phase, through the image detection the objects which can be grasped are recognized followed by the planning of trajectory motion of the gripper arm.
    - Then through the open or closed loop architecture the execution is achieved by the controllers and motors.
    - This paper focuses on defining the better grasp for the robot gripper. The authors used the five dimensional representation to detect the good grasp.
    - Although deep learning had been emerged substantially in computer vision and NLP, it had been limited in application into robotics.
    - So, the authors took the best use of it to implement and get the state-of-the-art results by deep convolutional neural networks.

3. **Related Works:**

- Deep learning had been used to train the movements of the robotic arm and localization purposes. The challenge of implementation of the deep learning is requirement of large datasets.
- Single window detection is the general technique used for 2-D robotic grasp pose detection. However, Lenz et al., defined a CNN architecture for the grasping which is very slow in prediction and execution.
- Research had been done in application of deep learning into 3-D object detection and pose recognition, which resulted in higher computational expense.
- Other works which had resulted in 92% of grasp detection had been limited to cloth towels and cannot be extended to general usage.
- Instead of application of AlexNet in prior works, the authors implemented the deep CNN ResNet in addition to the multi-modal feature extraction.

4. Approach and Experiments:

- Five dimensions, gripper length(h), width between two grippers(w), (x,y) position at center of the object and orientation, were introduced for every object. In general, h and w were put constant for the specific robots.
- The authors had adapted single step prediction rather than the two step prediction implemented in the prior works.
- As the depth is increased, the networks would result in more accuracy and better efficiency. It is also observed that SGD cannot be better optimizer for the deep neural networks as it would add up the training errors as it goes deep.
- So, Resnet is implemented in which the residual layers were mapped with certain functions in between.
- Some of the previous assumptions were also adapted, such as only a single graspable object is present. This assumption is favourable as it looks up the whole image and create the grasping position in the global coordinates.
- Replacing the AlexNet of 8 layers, the authors came up with fifty layer residual network for achieving the goal.
- Furthermore two architectures were introduced for the prediction as:
    - Unimodal grasp predictor: To detect the grasp pose, it uses only 2D image of RGB colors.
    - Multimodal grasp predictor: It uses 3D grasp prediction of RGB colors along with the depth information.
- In the unimodal architecture the ultimate two layers were replaced with ReLU replacing the ResNets. The ultimate fully connected layer will result in predicting the output.
- The multi-modal architecture as described above uses the multiple data to extract the grasping pose. However, even in this architecture the last two layers will only responsible for the prediction.

- Later a dropout layer is added to overcome the overfitting problem. With the help of two DCNN architectures working parallel, they extract the features of the image.
- The experiments were done on the standard dataset known as Cornell Grasp dataset, to evaluate and compare the results to prior works.
- Inheriting the previous implementation, the authors also took the help of five fold cross validation, in which they had splitted the dataset into two types as:
    - Image-wise split: Converts the image into random five folds. It helps in knowing the network's capability of generalizing objects, which the network had seen before.
    - Object-wise split: Splits the datasets based on objects and all the same objects were put into single set. It helps in knowing the network's capability of generalizing objects, which the network had not seen before.
- The pre processing is carried out before training and admitting data into the DCNN. In this process the pixels which have NaN values were substituted by zeros for calculation purposes.
- Pretraining is also done on the dataset as the considered dataset is domain-specific. Hence the Resnet was initially trained on Imagenet which classified the images into 1000 classes.
- The training had been done on the advanced hardware NVIDIA GTX 645 GPU with intel i7 processor. However it is not common to use such high processors in robots, it is stated that the new robots had been implementing with them due to increase in complex applications and computation power.

5. Results:

- Works had been evaluated by two performance measures:
    - Point grasp metric: It defines the error between the predicted center point of the object and the actual one.
    - Rectangle grasp metric: This evaluates the percentage coincidence of actual rectangle measure with the predicted rectangle with sides of gripper height(h) and width(w) between them.
- Both uni-modal and multi-modal architectures had outperformed the previous state of the art results with 88.53% and 89.21% respectively with Image-wise split.
- The replacement of ResNet-50 with pretrained VGG16 had also achieved good results better than the previous but less than the authors' ResNet results.
- The speed also increase by 800 times than the Lenz et al., two stage SAE model with 9.71 fps.
- It is also stated that for pretraining Imagenet may be replaced with other datasets, as the authors anticipated to achieve better results.

6. Conclusion:

- The paper had successfully achieved the state of the art results with new approach of implementing deep Convolutional Neural Networks into robotics.
- Despite of powerful hardware, it is obvious that the architecture can be easily adapted and implemented on different objects.
- Further the works can be concentrated on transfer learning on the already trained model to predict better grasp poses.
- Applications into industrial usage can also result in very good accuracies as the implementation of pick and place is very limited to specific objects.