Object localization and grasping area detection for social robots

Fupin Yao

Visiting student at IMI, NTU

Advisor: Prof. Nadia Thalmann

Overview

- Objective
- Challenge
- The state of the art
- Problem description
- Our Framework
- Experiments

Objective

-General problems of grasping

- What are the objects in front of them

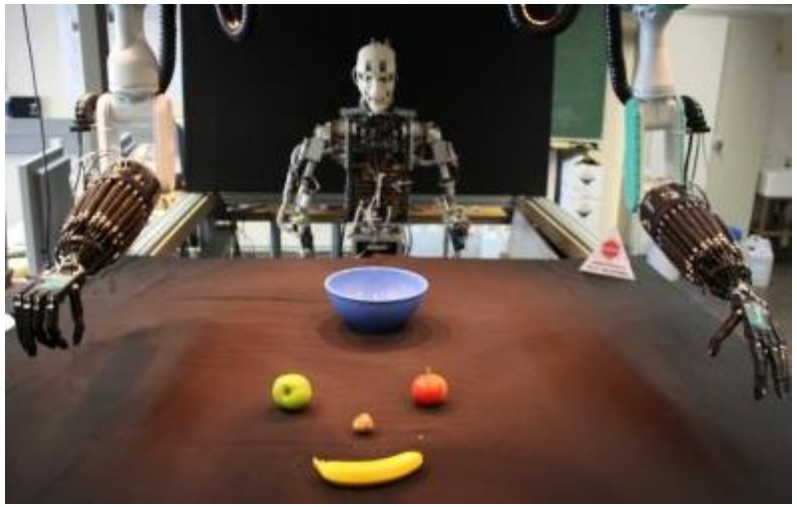
  Make sure what is the object

- where are the objects

  Get position of objects

- which part of objects should the robot grasp

  Get position of proper grasping area



Grasping[1]

Reference:
[1] Dornbusch, R.P., Haschke, D., Ritter, R. and HelgeContact (2012) Search. Available at: https://ni.www.techfak.uni-bielefeld.de/node/2914 (Accessed: 17 November 2016).

Challenge

- What and where are the objects?

  High accuracy in object recognition but low in detection

  Not real-time

Team name	Entry description	**Number of object categories won**	mean AP
CUImage	Ensemble of 6 models using provided data	109	0.662751
Hikvision	Ensemble A of 3 RPN and 6 FRCN models, mAP is 67 on val2	30	0.652704
Hikvision	Ensemble B of 3 RPN and 5 FRCN models, mean AP is 66.9, median AP is 69.3 on val2	18	0.652003
NUIST	submission_1	15	0.608752
NUIST	submission_2	9	0.607124
Trimps-Soushen	Ensemble 2	8	0.61816

Table 1: Object detection in ImageNet challenge 2016[1]

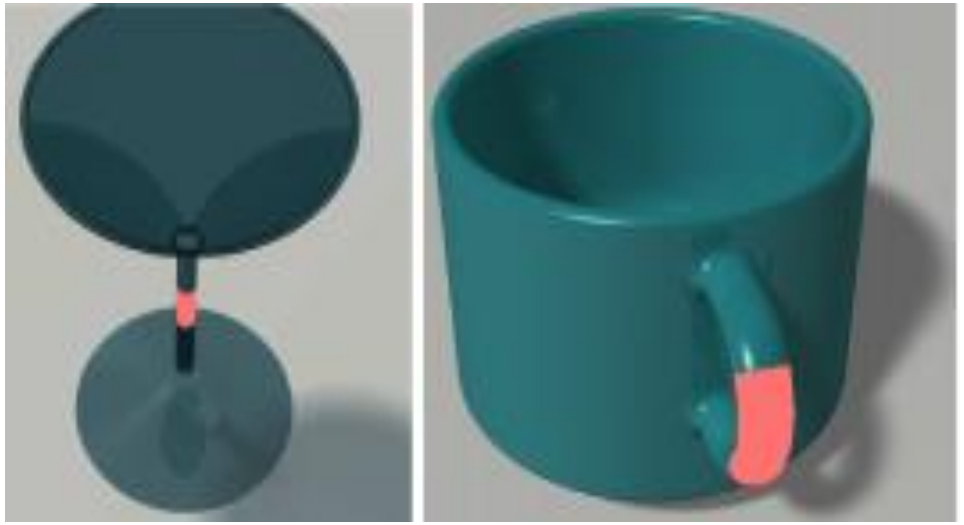
Reference:
[1] Lab, U.V. (2016) ILSVRC2016. Available at: http://image-net.org/challenges/LSVRC/2016/results (Accessed: 24 October 2016).

Challenge

- Which part should the robot grasp?

  Less researched

  Recent work:
  - –Saxena, Driemeyer, and Ng, 2008[1]
  - –Lenz, Lee, and Saxena, 2015[2]



Proper grasping parts [1]

Reference:
[1] Saxena A, Driemeyer J, Ng A Y. Robotic grasping of novel objects using vision[J]. The International Journal of Robotics Research, 2008, 27(2): 157-173.
[2] Lenz, I., Lee, H. and Saxena, A. (2015) 'Deep learning for detecting robotic grasps', The International Journal of Robotics Research, 34(4-5), pp. 705–724. doi: 10.1177/0278364914549607.

Challenge

- Grasping solution for social robots

  Less researched

  Hands more complicated

  Perform like a real human

ReFlex robotic hand [1]

Baxter robotic hand [2]

Real human hand [3]

Perform like a real human [4]

Reference:
[1] Introducing the reFlex hand (no date) Available at: http://docs.righthandrobotics.com/main:reflex (Accessed: 17 November 2016).
[2] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5):705-724, 2015.
[3] Home (2013) PngImg.Com. Available at: http://pngimg.com/download/878 (Accessed: 17 November 2016).
[4] Brain, S. (2016) Coffee drinking statistics. Available at: http://www.statisticbrain.com/coffee-drinking-statistics/ (Accessed: 17 November 2016).

The state of the art

-Object detection

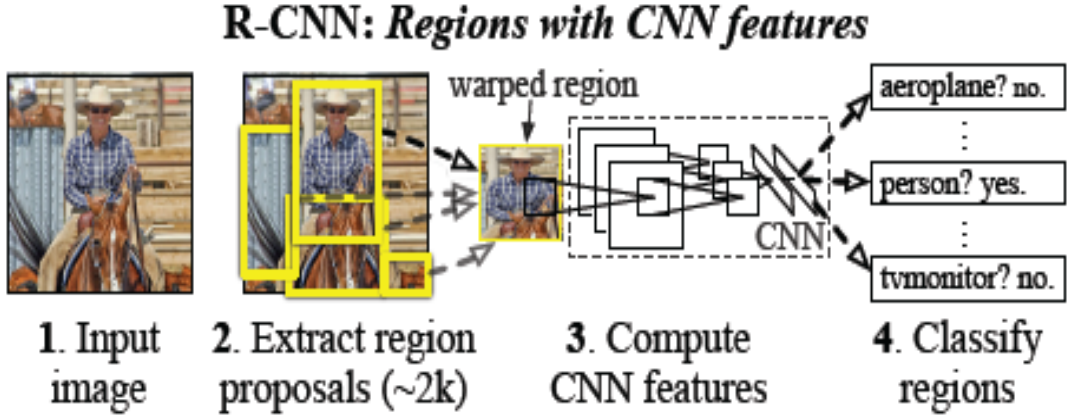
- R-CNN and fast R-CNN[1][2][3]

  Two steps:

  -Generating object proposals
  -Classifying proposals

  Limitation:

  -Repeated computation



R-CNN work flow [1]

Reference:
[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
[2] R. Girshick. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pages 1440-1448, 2015.
[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91- 99, 2015.

The state of the art

-Object detection

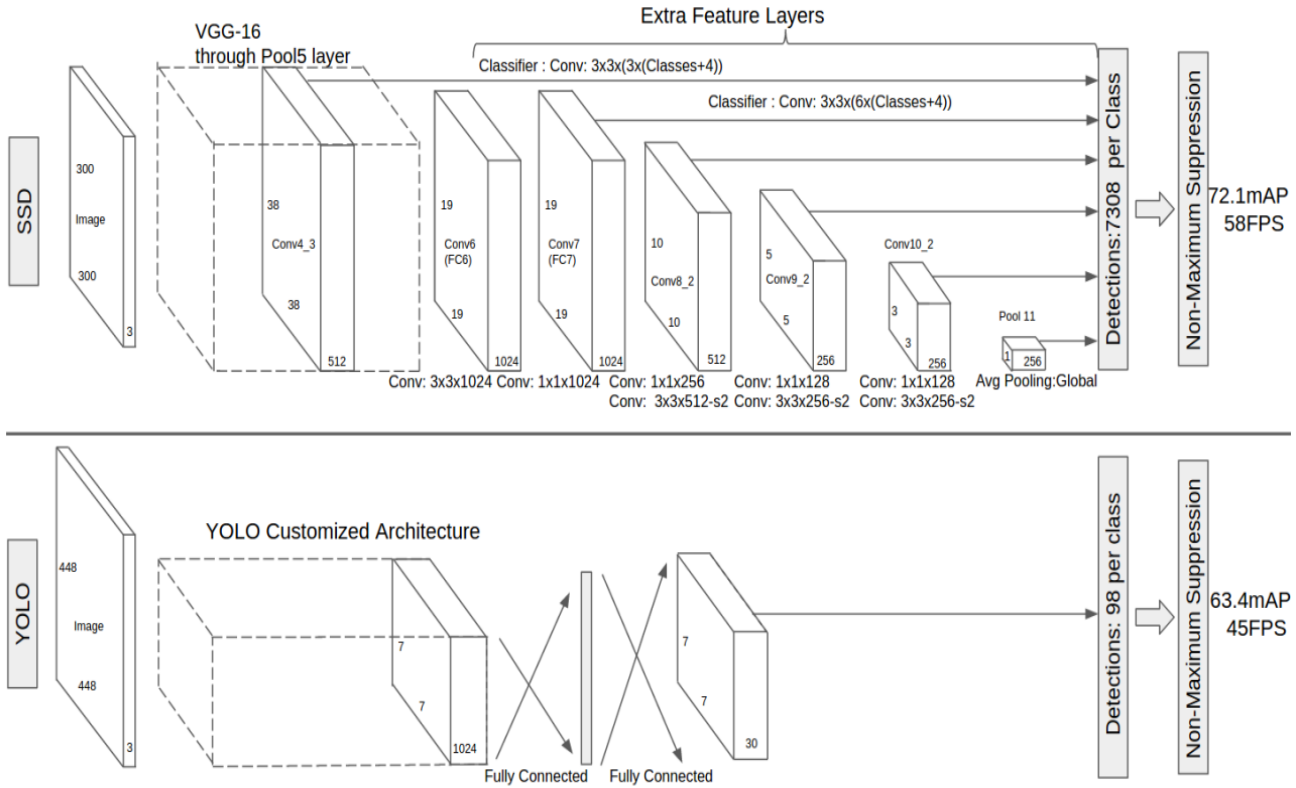
YOLO [1]

Eliminates bounding box proposals

Performs not well for small objects

SSD [2]

Uses multiple feature maps to perform detection at multiple scales



YOLO and SSD network architecture [2]

Reference:

[1] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. (2016)

[2] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. arXiv preprint arXiv:1512.02325, 2015.

The state of the art
-Grasping area detection

Robotic grasping of novel objects using vision[1]

First use visual information to predict grasping region

Learning to grasp objects with multiple contact points[2]

Get multiple grasping points

Common limitation

-Use local hand-designed features
-Time-consuming,detect in sliding windows

Robotic grasping of novel objects using vision [1]

Reference:
[1] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. The International Journal of Robotics Research, 27(2):157–173, 2008
[2] Quoc V Le, David Kamm, Arda F Kara, and Andrew Y Ng. Learning to grasp objects with multiple contact points. In IEEE International Conference on Robotics and Automation (ICRA), pages 5062-5069, 2010.

The state of the art
-Grasping area detection

Deep learning based grasping area detection

Approach	Feature	Limitation
Deep learning for detecting robotic grasps[1]	use convolutional networks	Time-consuming because of detection in sliding window
Real-time grasp detection using convolutional neural networks[2]	real-time	Not accurate for small objects
Object discovery and grasp detection with a shared convolutional neural network[3]	Detect objects and grasping areas	Only works when objects on a horizontal plane Outputs only one grasp

Reference:
[1] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5):705-724, 2015.
[2] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In IEEE International Conference on Robotics and Automation (ICRA), pages 1316-1322, 2015.
[3] Guo D, Kong T, Sun F, et al. Object discovery and grasp detection with a shared convolutional neural network[C]//Robotics and Automation (ICRA), 2016 IEEE International Conference on. IEEE, 2016: 2038-2043.

Problem description

Pinch as

Most widely used by humans

Pinch[1]

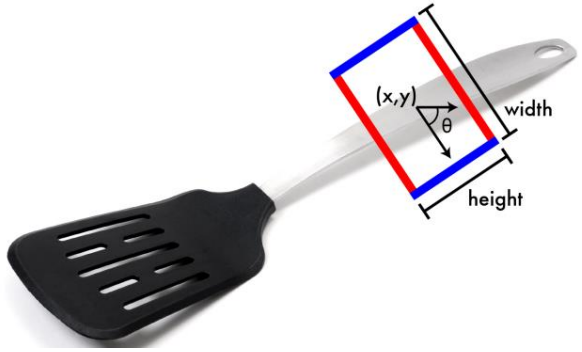
Pin representation[2]

g= $\{x, y, \theta, w, h\}$

x, y: center coordinates

θ: orientation angle

w, h: width and height



Pin representation[3]. The blue edges are parallel to pinching fingers and width is the open width of the two fingers

Reference:
[1] PINCH - Craighton Berman studio (2008) Available at: http://studio.craightonberman.com/PINCH (Accessed: 17 November 2016).
[2] [9] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5):705-724, 2015.
[3] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In IEEE International Conference on Robotics and Automation (ICRA), pages 1316-1322, 2015.

Problem description

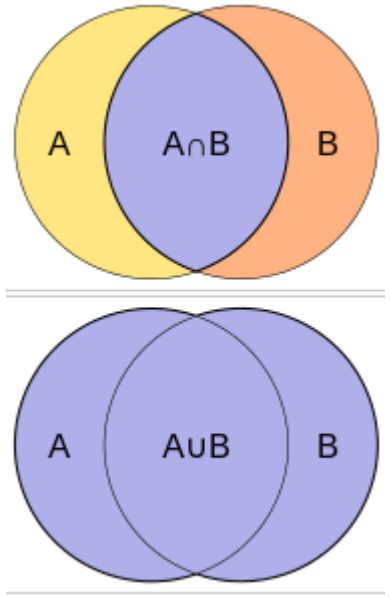
Grasping area metric[1][2]

Grasping area valid if:

-Angle difference less than 30 degrees

-Jaccard index greater than 0.25

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



Jaccard Index[3]

Reference:
[1] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5):705-724, 2015.
[2] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In IEEE International Conference on Robotics and Automation (ICRA), pages 1316-1322, 2015.
[3] Jaccard index (2016) in Wikipedia. Available at: https://en.wikipedia.org/wiki/Jaccard_index (Accessed: 17 November 2016).

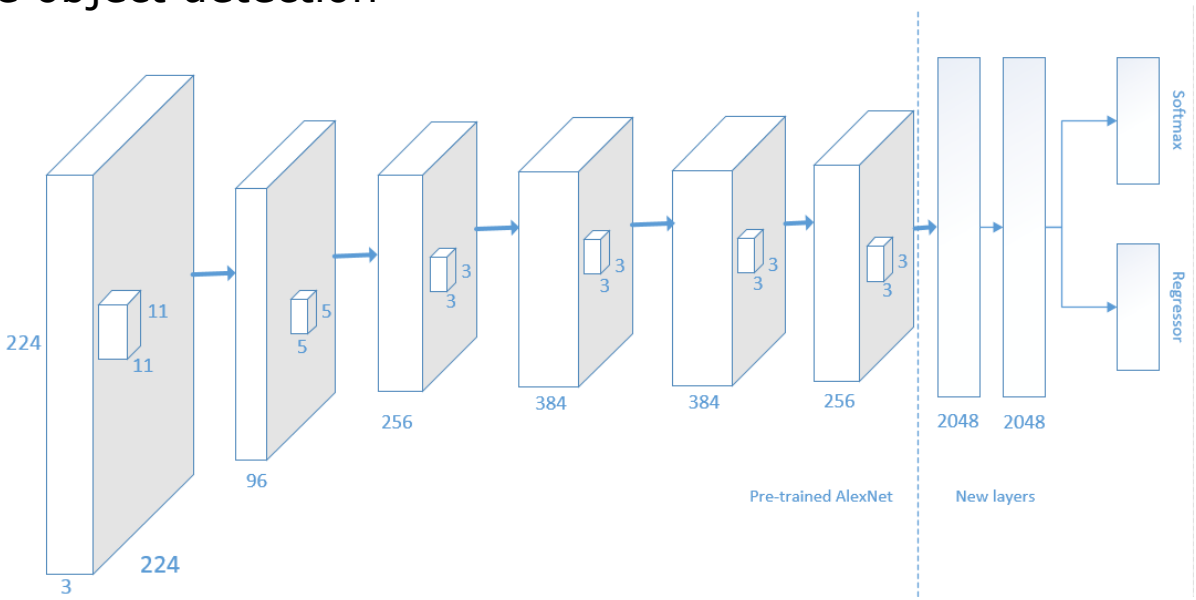
Oject localization and grasping area detection

Model Architecture

Pre-trained AlexNet[1] and new layers

Shared network architecture for object recgnition and grasping area detection

Not require object detection



Model architecture

Reference:
[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

Experiments(On going)

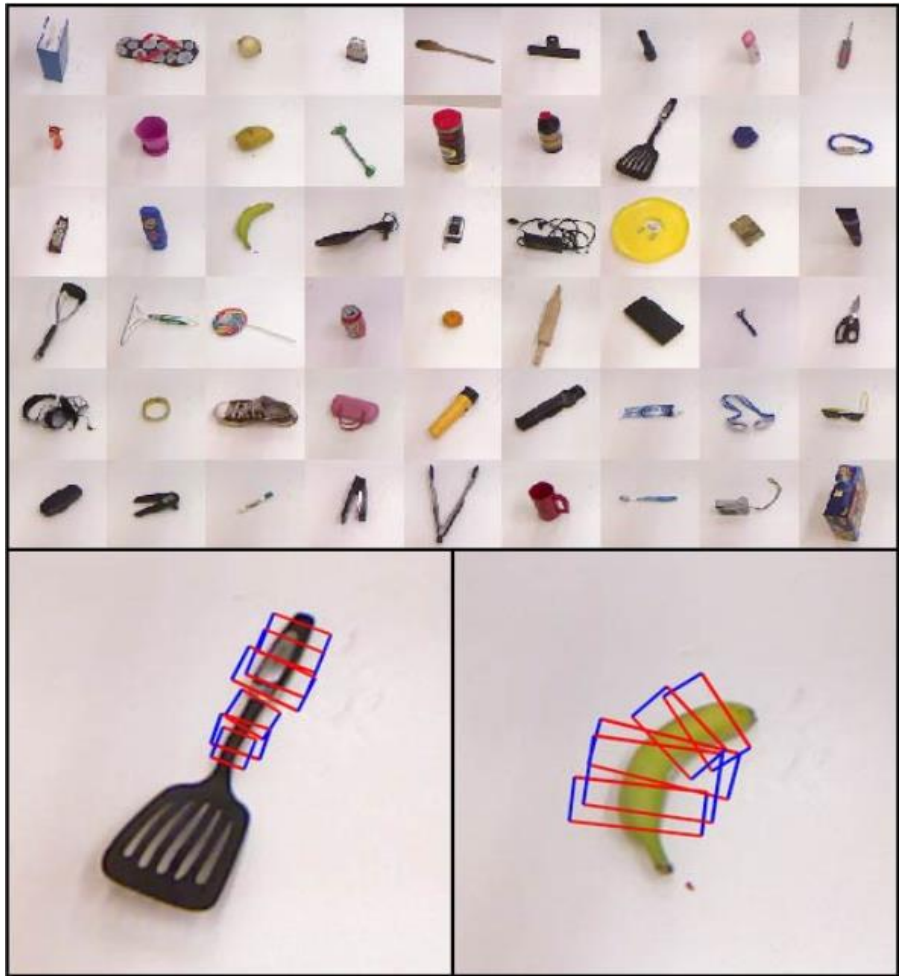
Data preprocessing

Cornell Grasping Dataset[1]: 885 images of 240 different object

Center crop of 340*340

Substitute blue channel with depth channel

Data augmentation: translate and rotate



Cornell Grasping Dataset[1]

Reference:
[1] Cornell (2009) Personal robotics: Grasping. Available at: http://pr.cs.cornell.edu/grasping/rect_data/data.php (Accessed: 18 November 2016).

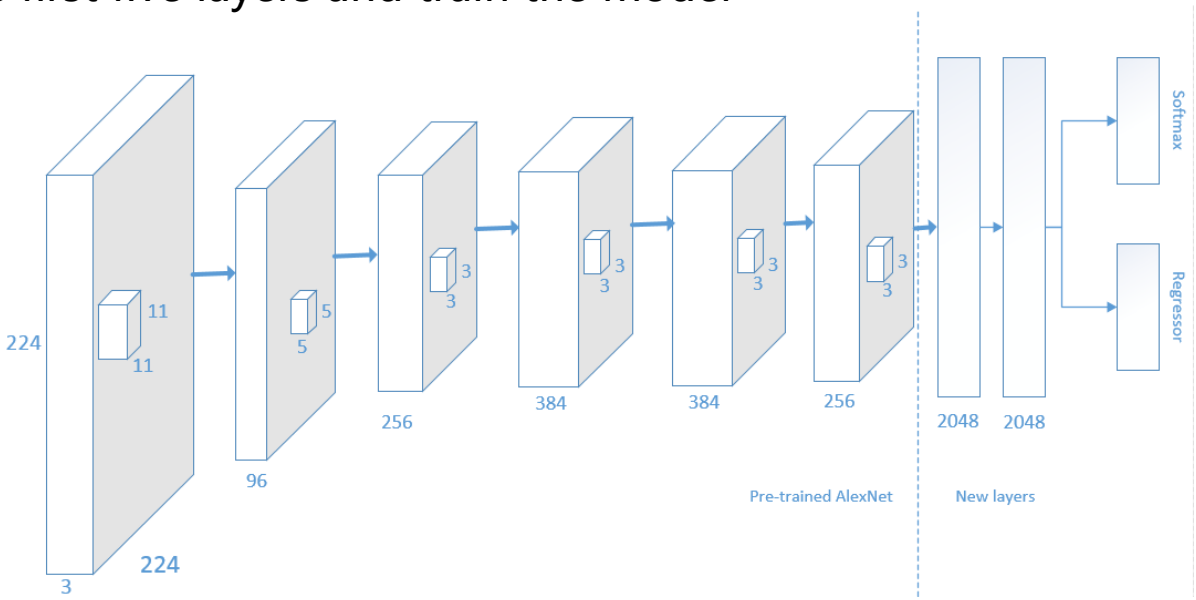
Experiments(On going)

Training

Fine-tune the model:

-Initialize the first five layers with AlexNet[1] weights

-Freeze the first five layers and train the model



Model architecture

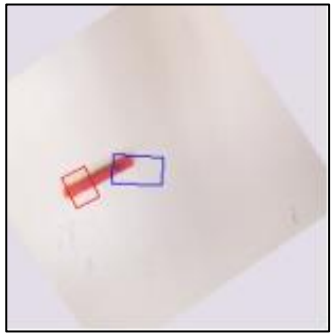
Reference:
[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

Experiments(On going)

Results:

Could predict location of objects

Orientation of grasping area not accurate



Software development

caffe: faster than theano

Hareware platform

Tian with 6GB memory, 500 GB disk

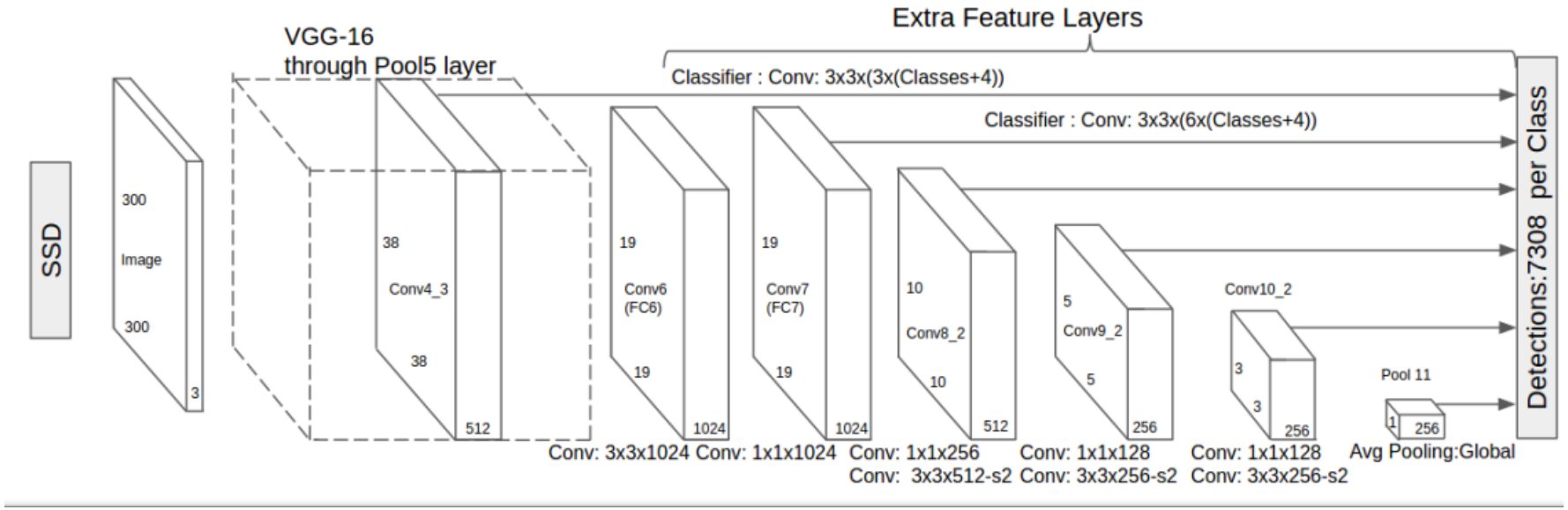
Future work

Solution for multi-grasp detection

-Add layers of different size to predict at multiple scales

-5 outputs for grasping area, c class scores

-derived from SSD[1]



Model architecture

Reference:

[1] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. arXiv preprint arXiv:1512.02325, 2015.

Thank you!

Q & A