Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Supersizing Self-supervision:
Learning to Grasp from 50K Tries and 700 Robot Hours

by Carl v. Heyden – 1.12.2022

Pinto, L. and Gupta, A. (2015). Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In 2015 IEEE international conference on robotics and automation (ICRA), pages 3406–3413. IEEE.

1. Introduction
2. Motivation
3. Related Work
4. Approach
5. Result
6. Conclusion
7. What came next?

Introduction

Supersizing Self-Supervision (2015):

Historical classification

- Published in 2015
- Has been cited 1048 times
- Many scientists referred to this paper and developed ideas further

How do we predict grasp locations for an object?

https://www.youtube.com/watch?v=oSqHc0nLkm8

How do we predict grasp locations for an object?

Modelling the object?
- Does not scale
- Ignores density
- Ignores mass distribution

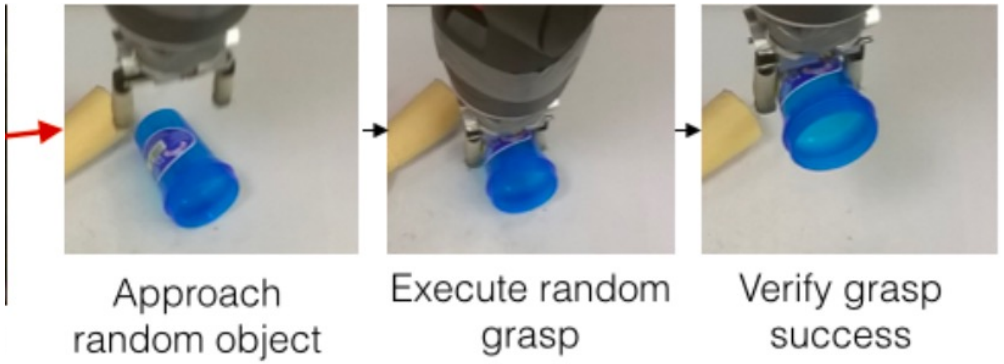
Annotating grasp positions on the mesh?
- Main development after release of the paper
- Easy for particular gripper

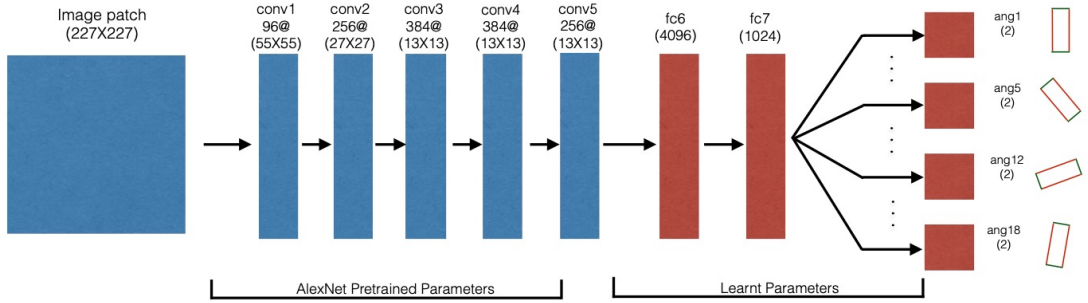
Connect the mesh with the real-world object using camera image?
- Different but active research field

Machine learning?
- Exhaustive human labelling impossible
- Biased by semantics

Train the robot in simulation?
- Different research field

Self-supervised grasping





Pinto, L. and Gupta, A. (2015)

- First approach to build a self-supervised setup and collect huge amount of data to train a neuronal network
- Trial and error

Goal

- Large dataset for the task of grasping
- Novel formulation of convolutional neuronal networks (CNN)
- Multi-stage learning approach

Approach

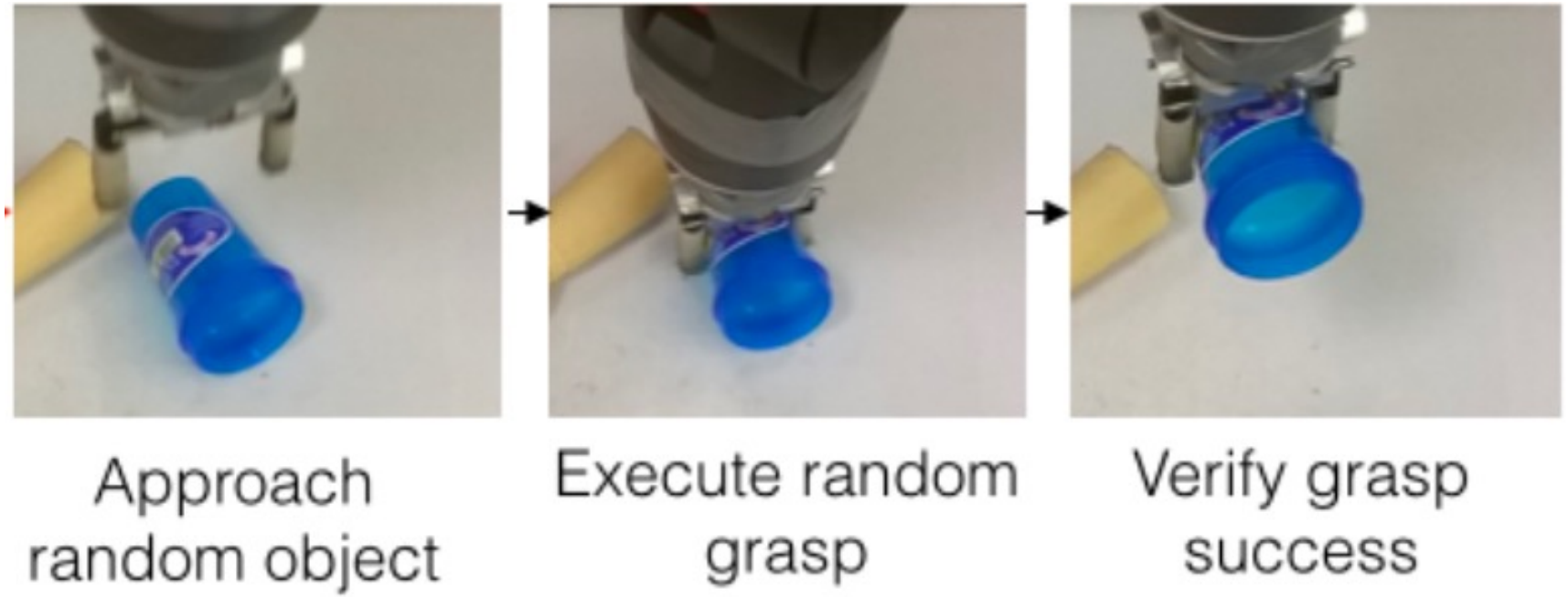
Robot grasping system:

- Baxter robot (both arms are used in parallel)
- ROS as development system
- 2 finger gripper
- Gripper force sensor
- Kinect V2 for table-top
- Small camera for end effector

Pinto, L. and Gupta, A. (2015)

Trial and error experiments:

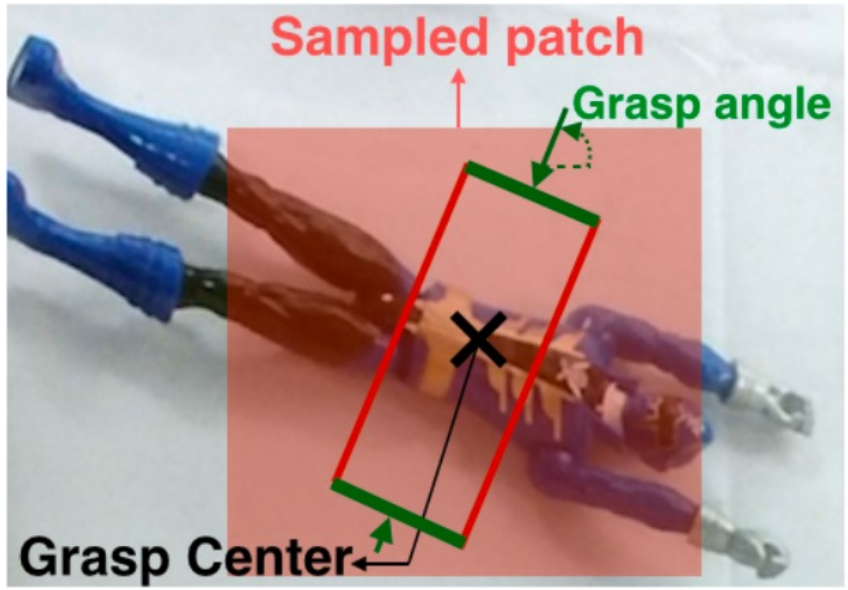
- Random grasp is executed and stored as failure or success. This is recognized by the gripper´s force sensor
- Images of table-top and end effector, robot arm trajectories and gripping history are recorded to disk



Pinto, L. and Gupta, A. (2015)

Problem formulation:

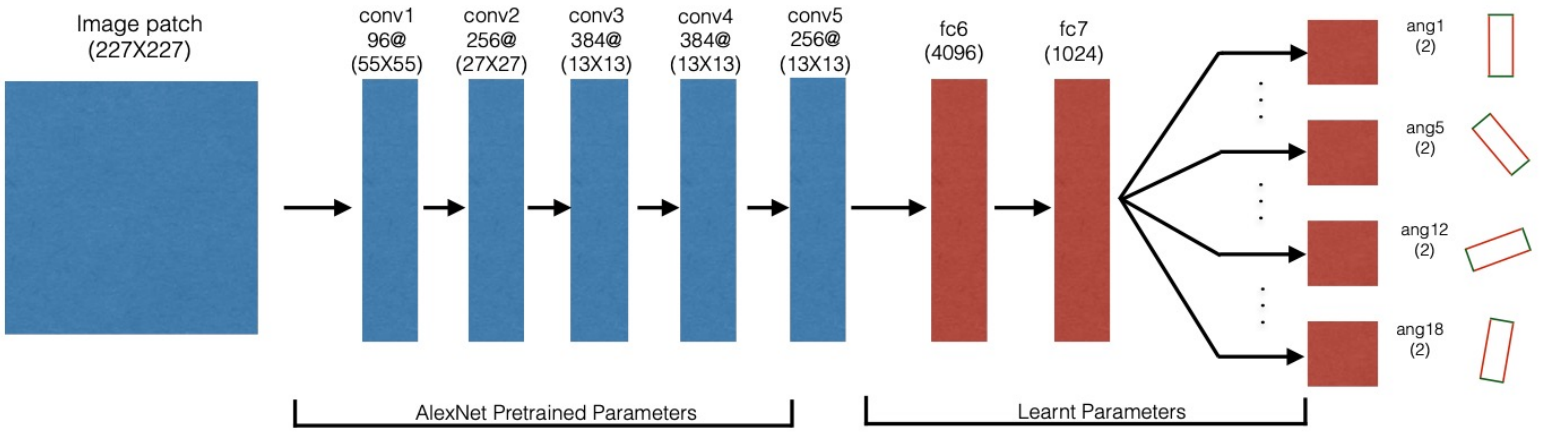
- Finding a successful grasp configuration of an object **I** using convolutional neuronal networks (CNN).
- AlexNet CNN model pre-trained on ImageNet is used
- Input to CNN
  - Image patch 1.5 times the gripper fingertips
- Output
  - Gripper orientation



Pinto, L. and Gupta, A. (2015)

Training approach:

- Data preparation
- Network design
  - AlexNet pretrained on ImageNet



Pinto, L. and Gupta, A. (2015)

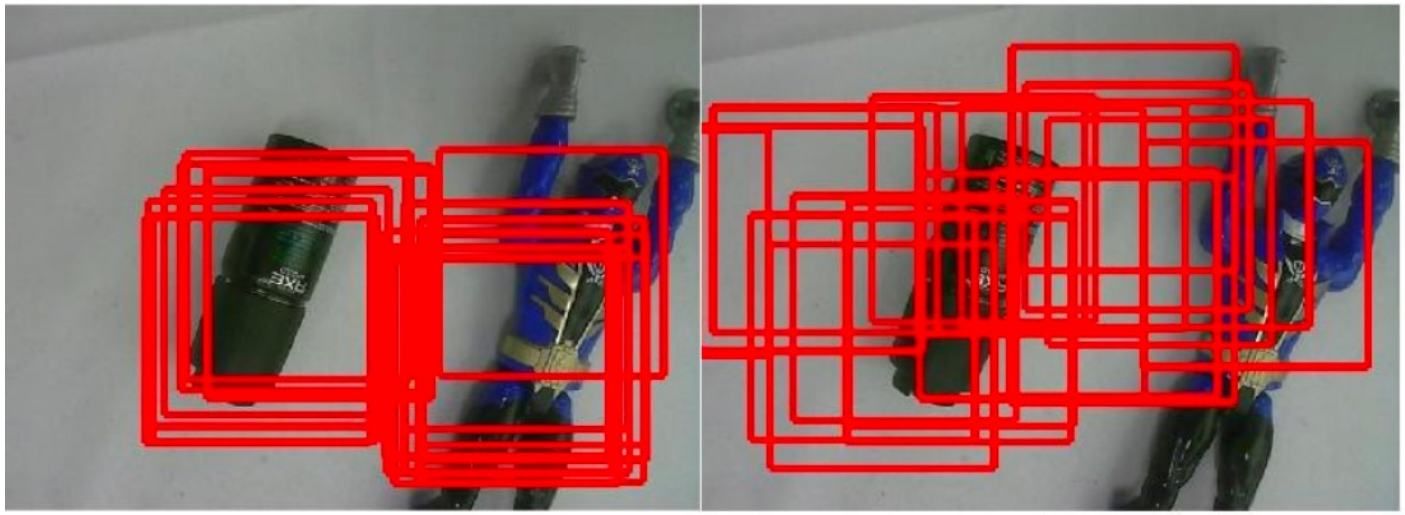
- Loss function

$$L_B = \sum_{i=1}^{B} \sum_{j=1}^{N=18} \delta(j, \theta_i) \cdot \text{softmax}(A_{ji}, l_i)$$

Staged learning:

Grasp selection:

- The robot now uses this trained network as the default for gripping.
- Seen and novel objects are used to enrich the model and avoids overfitting
- Procedure staged learning:
  1. 800 randomly sampled patches are evaluated by the trained network
  2. 800 x 18 grasp-ability prior matrix is generated
  3. Grasp execution is decided by importance sampling



a b

Pinto, L. and Gupta, A. (2015)

Staged learning:

Data Aggregation:
CNN training iteration: ***k***
Random grasp dataset for training: ***D***
The CNN uses the result from the last network train iteration ***k***-1 to finetune the network using the dataset ***D***.
Iteration 0 is simply trained on D

- Iteration 0: Learning rate 0.01 and 20 epochs
- Iteration > 0: Learning rate 0.001 and 5 epochs

Result

Training dataset:

- 150 Objects with varying graspability
- Cluttered table
- 50K grasp experience interactions

Data Collection Type	Positive	Negative	Total	Grasp Rate
Random Trials	3,245	37,042	40,287	8.05%
Multi-Staged	2,807	4,500	7,307	38.41%
Test Set	214	2,759	2,973	7.19%
	6,266	**44,301**	**50,567**	

Pinto, L. and Gupta, A. (2015)

Pinto, L. and Gupta, A. (2015)

Testing and evaluation setting:

- Test on objects not seen in the training
- 3000 physical robot interactions on 15 new objects in multi poses
- Binary evaluation: grasped or not grasped
- Accuracy of 79.5% on this test set

Comparison with baselines:

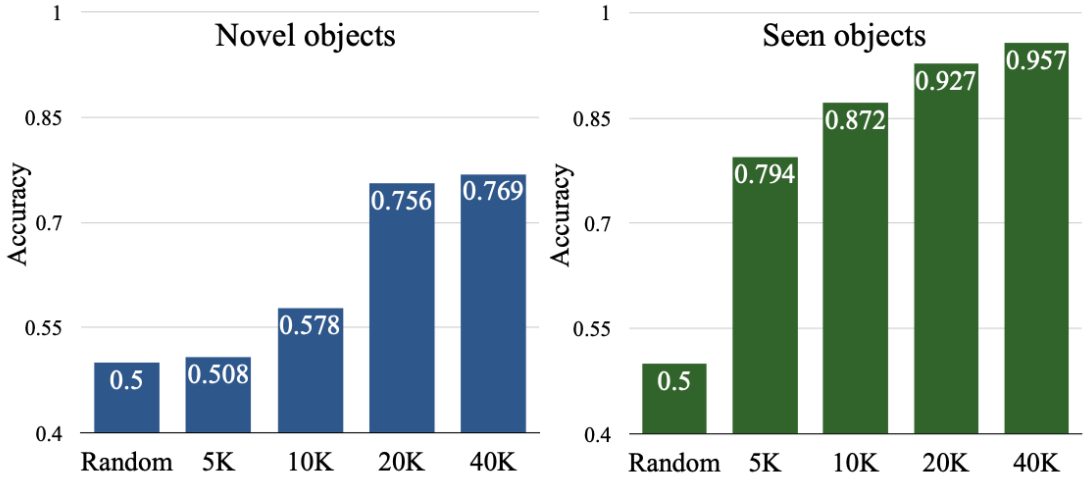
COMPARING OUR METHOD WITH BASELINES

	Heuristic			Learning based			
	Min eigenvalue	Eigenvalue limit	Optimistic param. select	kNN	SVM	Deep Net (ours)	Deep Net + Multi-stage (ours)
Accuracy	0.534	0.599	0.621	0.694	0.733	0.769	**0.795**

Pinto, L. and Gupta, A. (2015)

Ablative analysis

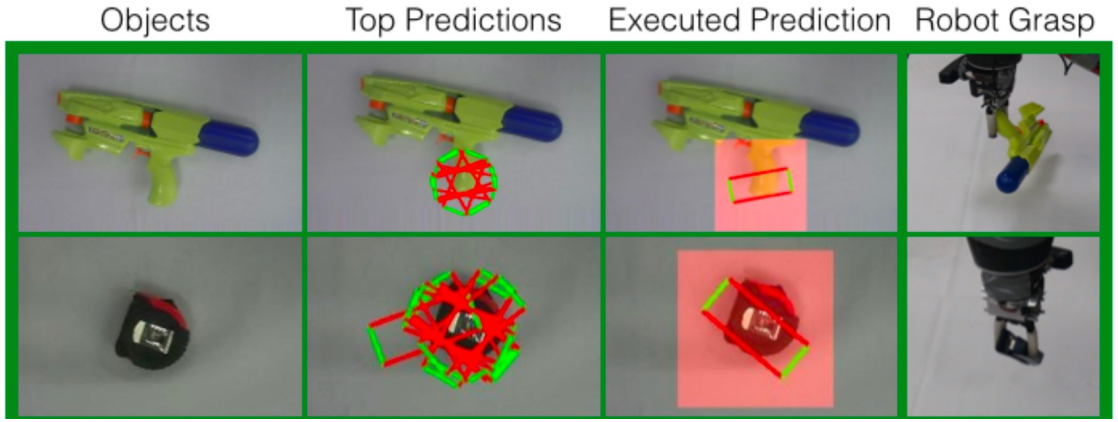
- Effects of data
  - More data helps to increase accuracy
- Effects of pretraining
  - Increase accuracy from 64.6% to 76.9%
- Effects of multi-staged learning
  - Increase accuracy from 76.9% to 79.5%
- Effects of data aggregation
  - Increase accuracy from 76.9% to 72.3%



Pinto, L. and Gupta, A. (2015)

Robot testing results

- Re-ranking grasps
  - Top 10 grasps on an object are identified
  - Neighbourhood analysis for every grasp **P**:
    - sample 10 patches in the neighbourhood of **P**
    - Get the best angle score of every patch and calculate the average angle score of all 10 patches
    - the calculated average is the new angle score of **P**
  - Re-rank the Top 10 grasps according to the new angle scores
  - **P** with the highest angle score after the neigbourhood analysis is executed



Pinto, L. and Gupta, A. (2015)

Conclusion

Conclusion

- First approach to build a self-supervised setup and collect huge amount of data
- Parallel execution approach accelerates data collection
- Pretraining, multi-stage learning and data aggregation increase the accuracy
- Reranking grasps minimizes errors

What came next?

Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection (Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D, 2016)

https://www.youtube.com/watch?v=V05SuCSRAtg

Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection (Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D, 2016)

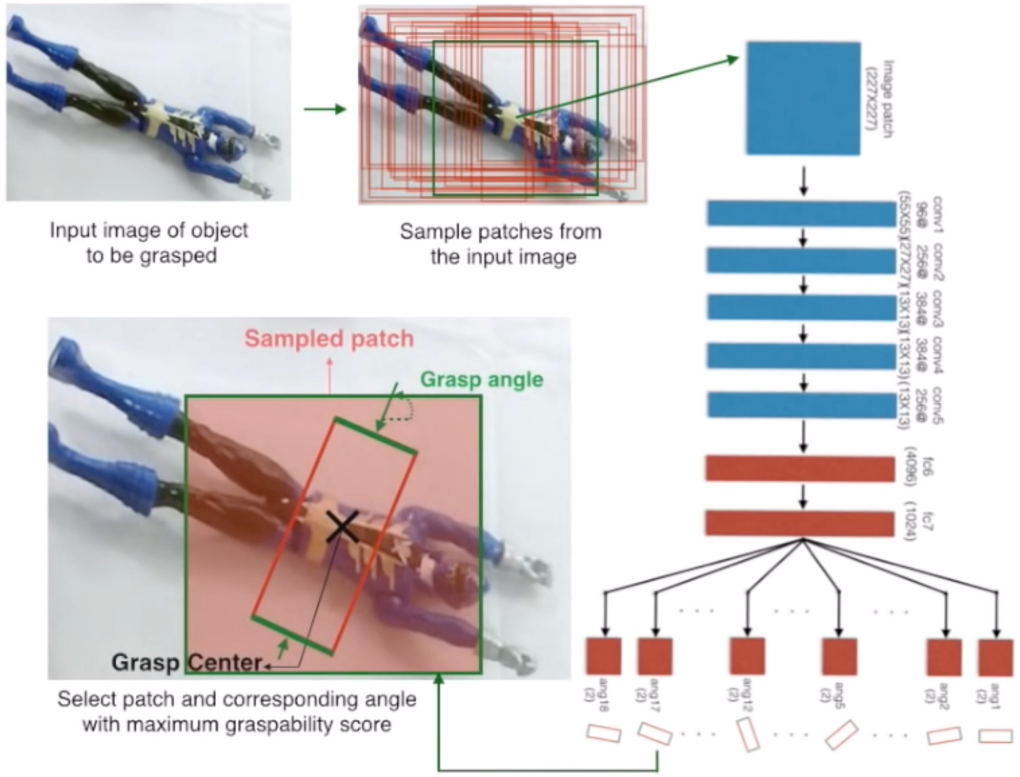
Data collection:

- Data collection with 14 robotic manipulators
- 800.000 grasps attempts

Complexity:

- Instead of predicting a grasp angle (parameter) the manipulators learn an actual policy which results in real-time control of the gripper using a camera

Thank you for your attention



Questions?

https://www.youtube.com/watch?v=UlUqTaYNER8&feature=emb_title