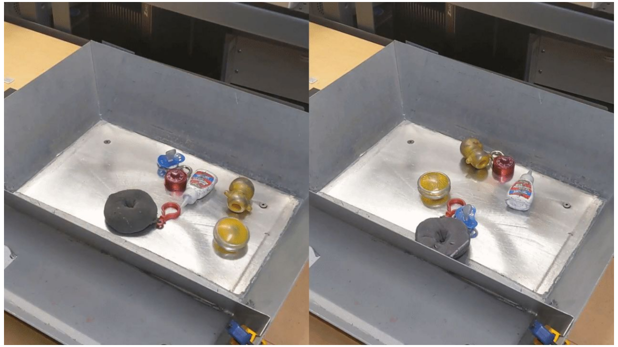
Never Stop Learning: The Effectiveness of Fine-Tuning in Robotic Reinforcement Learning



Ryan Julian
November 18th, 2020

Ryan Julian, Benjamin Swanson, Gaurav S. Sukhatme, Sergey Levine, Chelsea Finn, Karol Hausman

Website: https://ryanjulian.me/never-stop-learning

Google AI

USC University of Southern California

Roadmap

- **Problem**
- Preliminaries
- Baseline Study
- Fine-Tuning for Off-Policy RL
- A Very Simple Fine-Tuning Method
- From Fine-Tuning to Continual Learning
- Insights and Issues

Problem: How to make robots (continually) adapt?

End-to-end RL: Lots of success, but mostly it looks a lot like supervised learning

1. **Collect** (a bunch of) data
2. **Learn** from that data
3. **Deploy** learned model
4. (there is no 4th step)

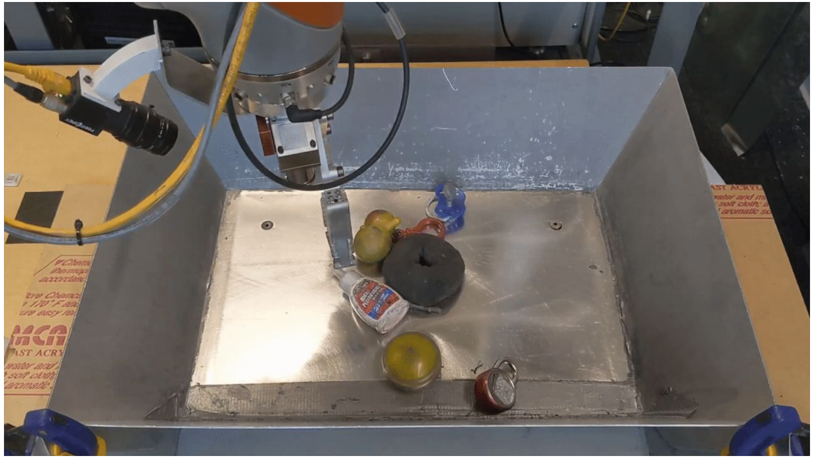
The **promise** of RL:

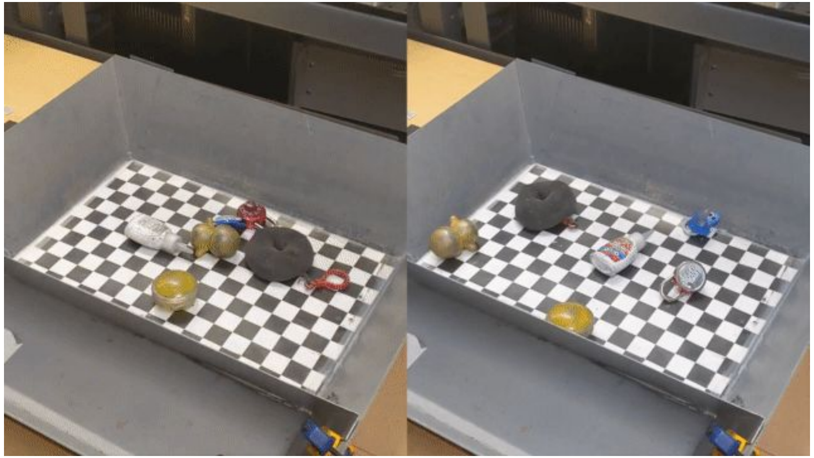
1. **Collect** data
2. **Learn**
3. **Deploy**
4. **GOTO** 1

Problem: How to make robots (continually) adapt?

94%

50% ⟶ 90%

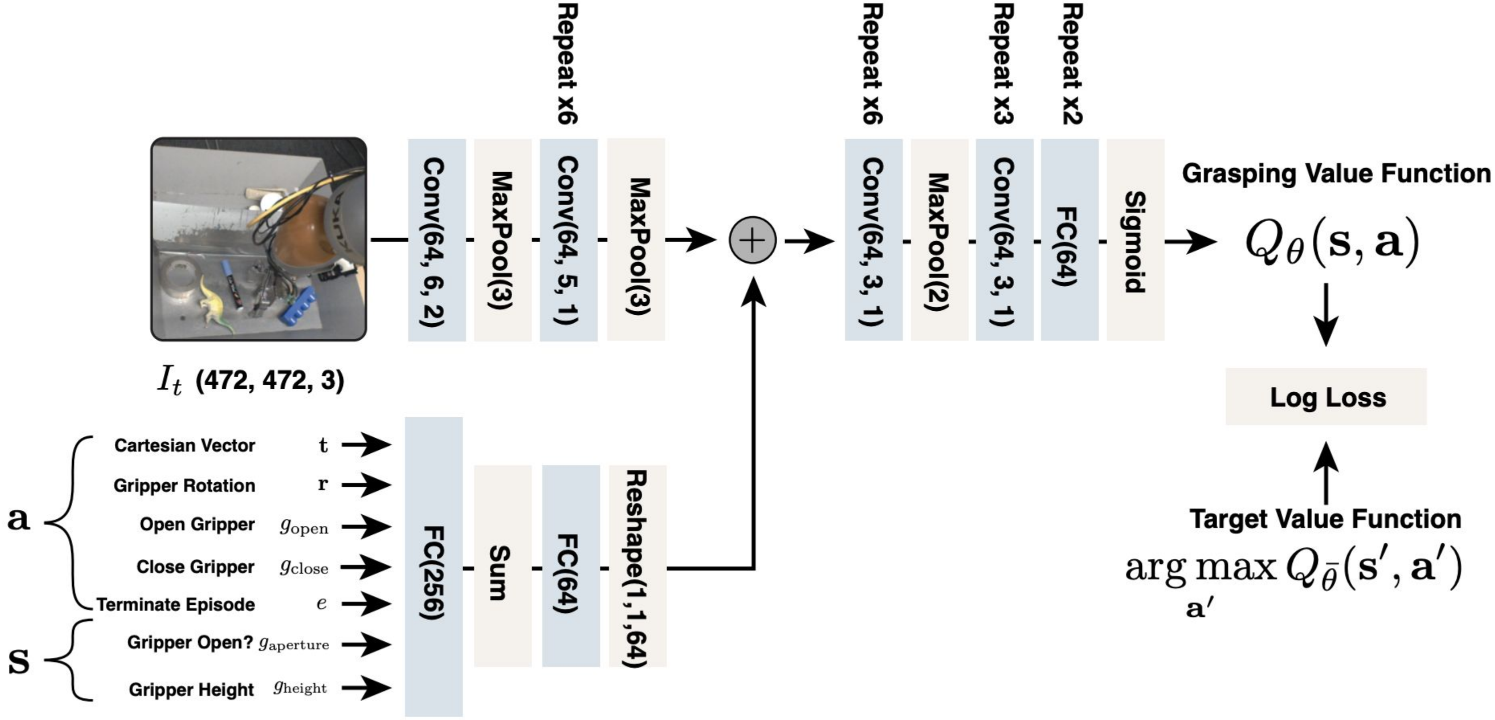




Roadmap

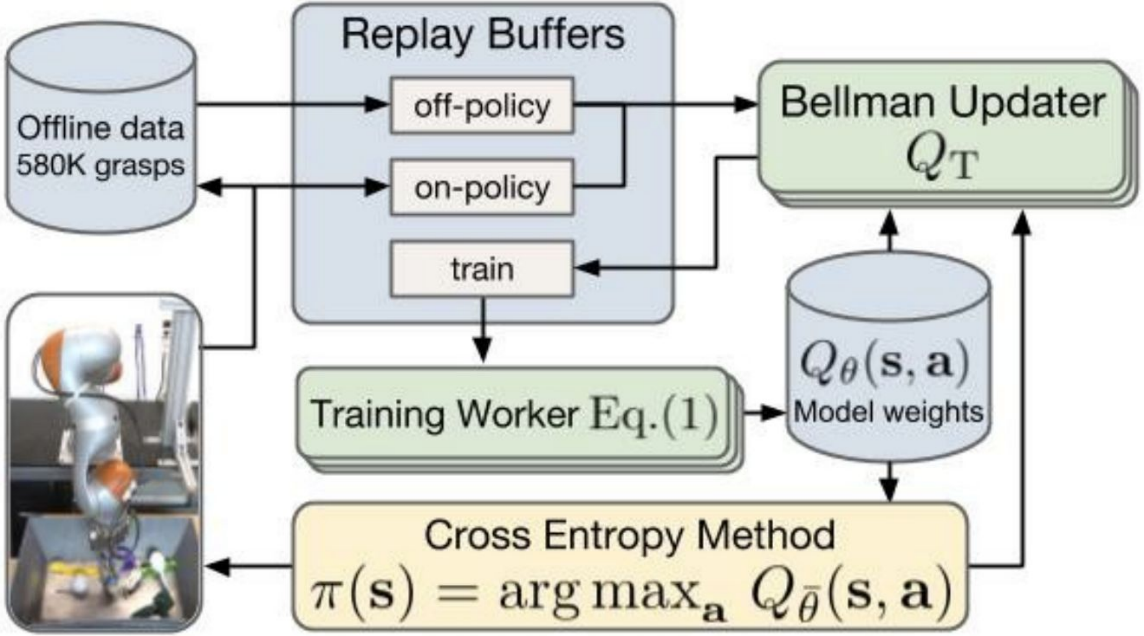
- Problem
- **Preliminaries**
- Baseline Study
- Fine-Tuning for Off-Policy RL
- A Very Simple Fine-Tuning Method
- From Fine-Tuning to Continual Learning
- Insights and Issues

Preliminaries: QT-Opt Grasping Architecture

- 

Source: QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation. Kalashnikov, et al. 2018.

Preliminaries: QT-Opt

- 

Source: QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation. Kalashnikov, et al. 2018.

Roadmap

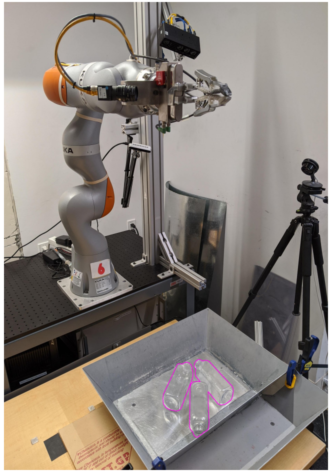
- Problem
- Preliminaries
- **Baseline Study**
- Fine-Tuning for Off-Policy RL
- A Very Simple Fine-Tuning Method
- From Fine-Tuning to Continual Learning
- Insights and Issues

Baseline: Robustness of Visual Grasping Policies

- Visual end-to-end RL is surprisingly robust
- No change: most backgrounds, most new objects, broken gripper, normal lighting, offset gripper by up to 5cm

Baseline: Robustness of Visual Grasping Policies

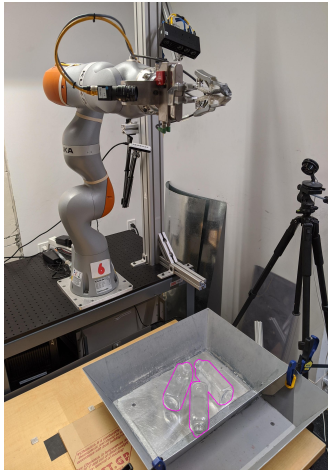
- Visual end-to-end RL is surprisingly robust
- No change: most backgrounds, most new objects, broken gripper, normal lighting, offset gripper by up to 5cm



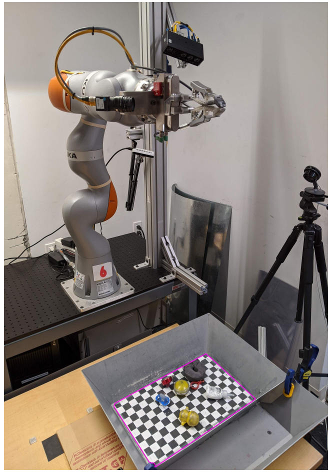
Transparent
Bottles

Baseline: Robustness of Visual Grasping Policies

- Visual end-to-end RL is surprisingly robust
- No change: most backgrounds, most new objects, broken gripper, normal lighting, offset gripper by up to 5cm



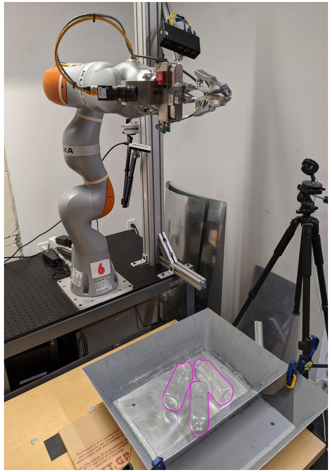
Transparent Bottles



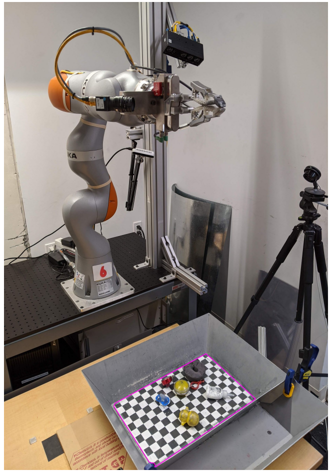
Checkerboard Backing

Baseline: Robustness of Visual Grasping Policies

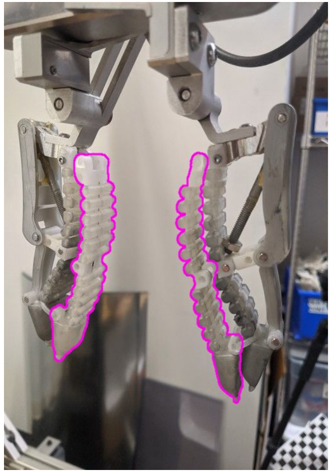
- Visual end-to-end RL is surprisingly robust
- No change: most backgrounds, most new objects, broken gripper, normal lighting, offset gripper by up to 5cm



Transparent Bottles



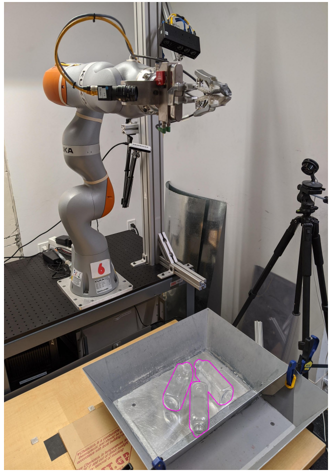
Checkerboard Backing



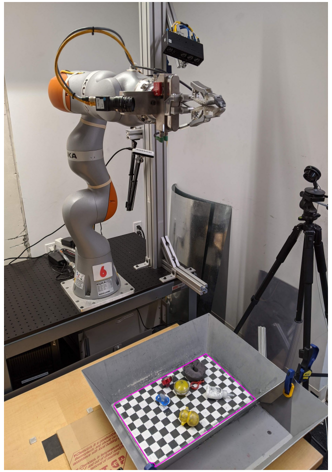
Extend Gripper 1cm

Baseline: Robustness of Visual Grasping Policies

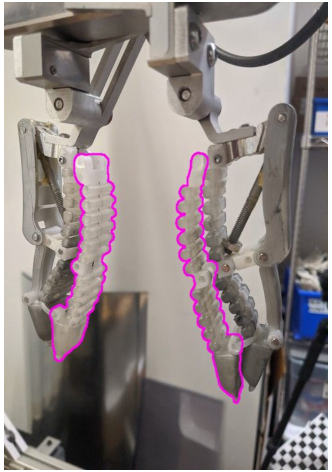
- Visual end-to-end RL is surprisingly robust
- No change: most backgrounds, most new objects, broken gripper, normal lighting, offset gripper by up to 5cm



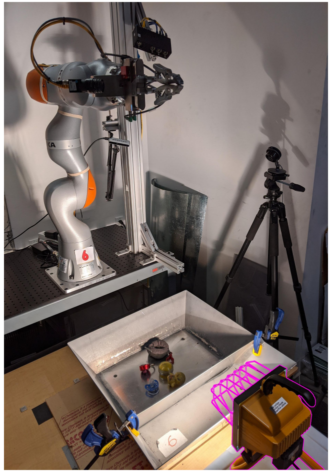
Transparent Bottles



Checkerboard Backing



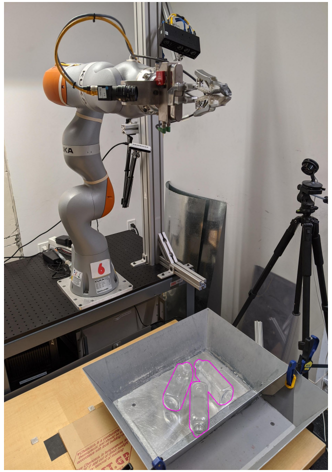
Extend Gripper 1cm



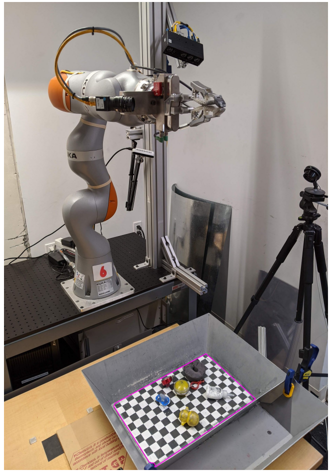
Harsh Lighting

Baseline: Robustness of Visual Grasping Policies

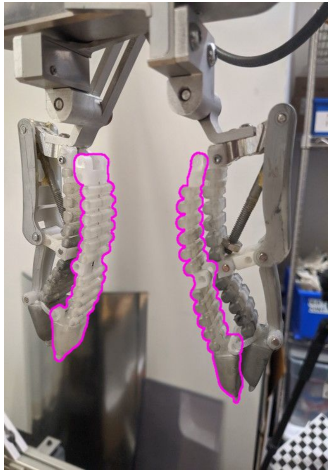
- Visual end-to-end RL is surprisingly robust
- No change: most backgrounds, most new objects, broken gripper, normal lighting, offset gripper by up to 5cm



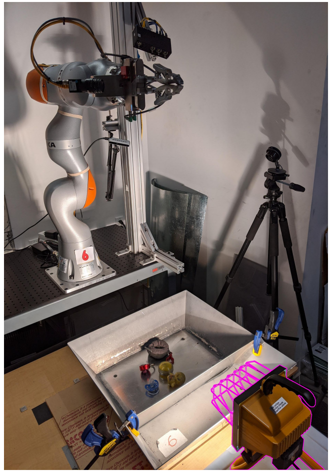
Transparent Bottles



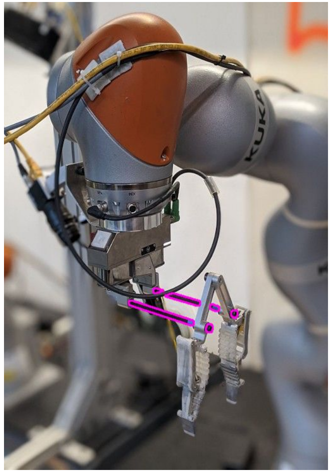
Checkerboard Backing



Extend Gripper 1cm

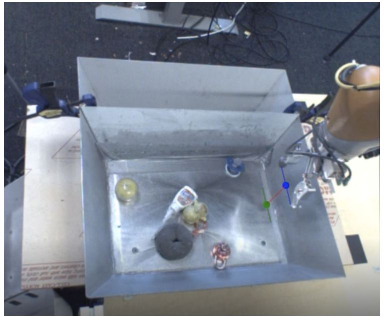


Harsh Lighting



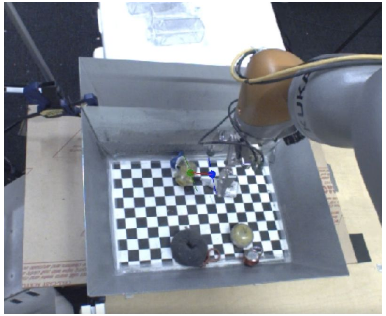
Offset Gripper 10cm

Baseline: What the robot sees

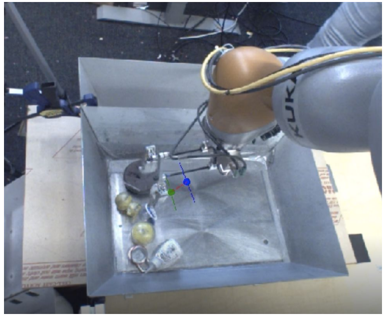


Base Grasping

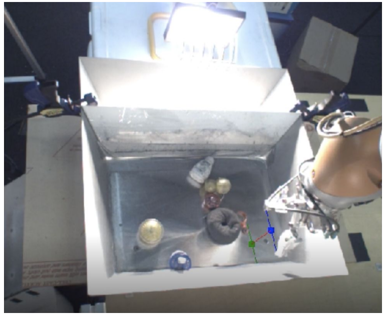
Extend Gripper 1cm



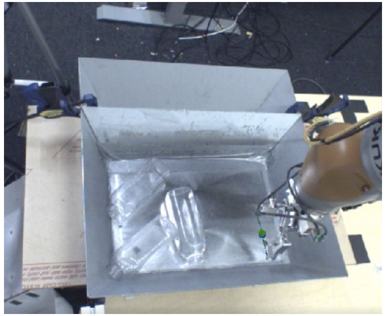
Checkerboard Backing



Offset Gripper 10cm



Harsh Lighting

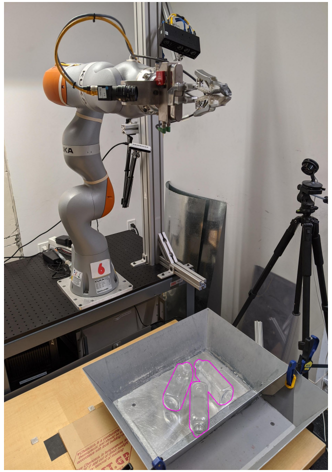


Transparent Bottles

Baseline: Robustness of Visual Grasping Policies

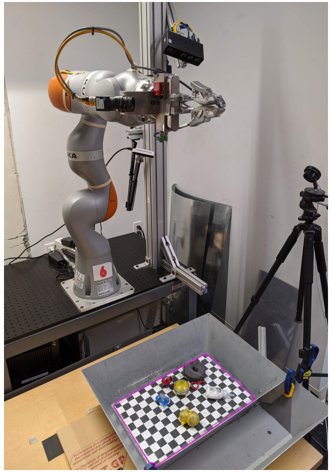
- Baseline study creates 5 challenge tasks

49%



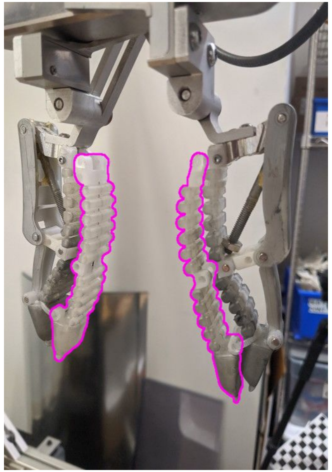
Transparent Bottles

50%



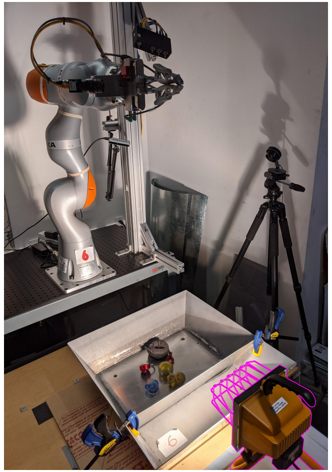
Checkerboard Backing

75%



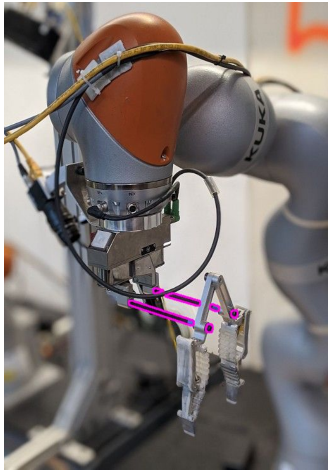
Extend Gripper 1cm

32%



Harsh Lighting

43%



Offset Gripper 10cm

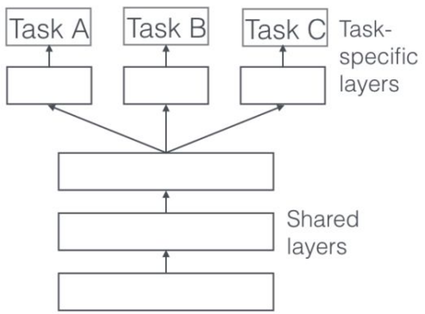
Roadmap

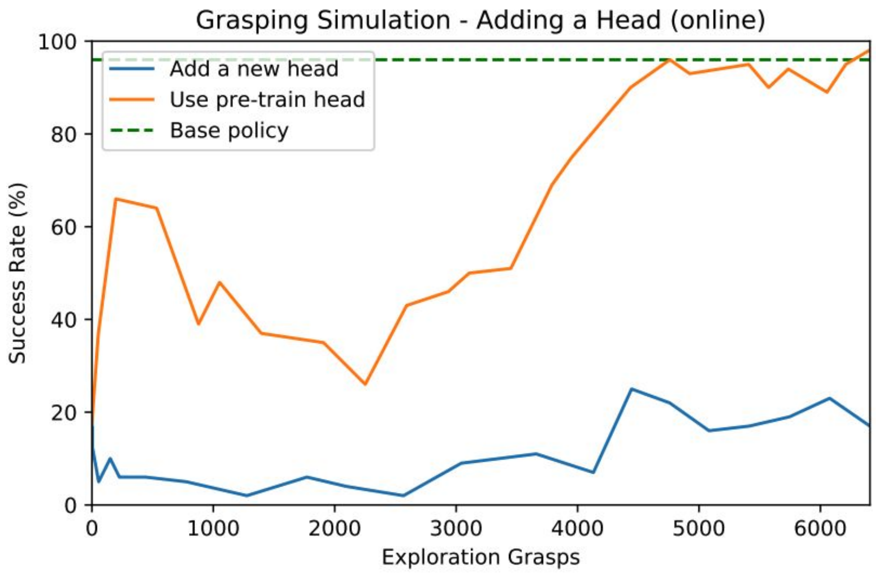
- Problem
- Preliminaries
- Baseline Study
- **Fine-Tuning for Off-Policy RL**
- A Very Simple Fine-Tuning Method
- From Fine-Tuning to Continual Learning
- Insights and Issues

Fine-Tuning for Off-Policy RL (vs. Supervised)

Case Study: Adding a "Head"

- **Conventional SL approach:**
  - Train the "body" + "head A" on base task
  - Discard "head 1", graft "head 2" onto network
  - Freeze "body" (or not), update network



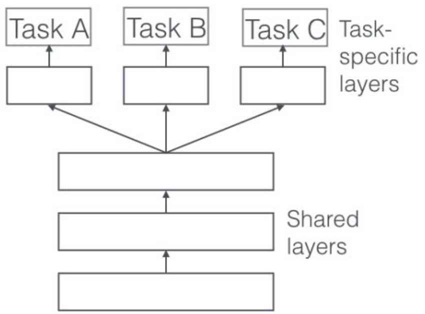


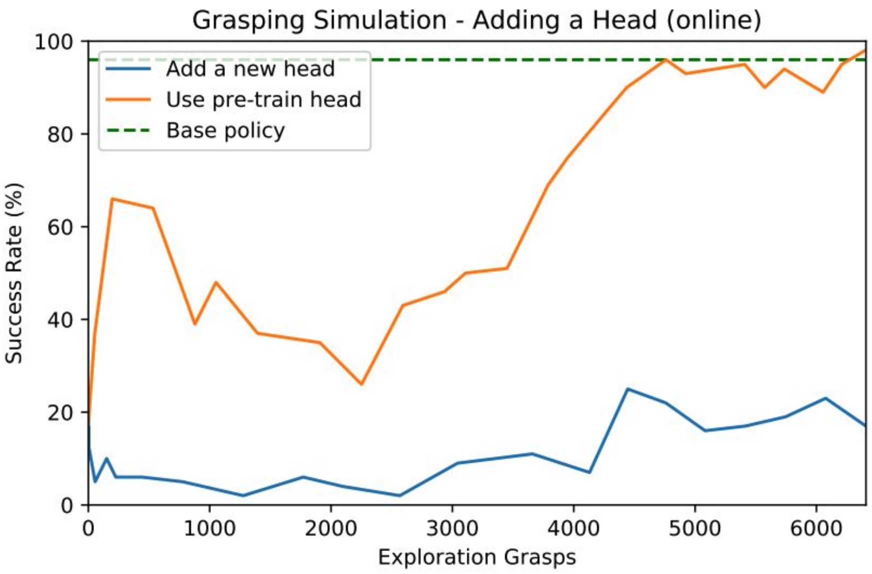
Source: An Overview of Multi-Task Learning in Deep Neural Networks. Ruder, 2017.

Fine-Tuning for Off-Policy RL (vs. Supervised)

Case Study: Adding a "Head"

- **Problem:** RL needs to explore
  - New head is uninformative for exploration
  - RL agent is unable to collect useful data for the new task
  - Same logic applies to other architectural approaches



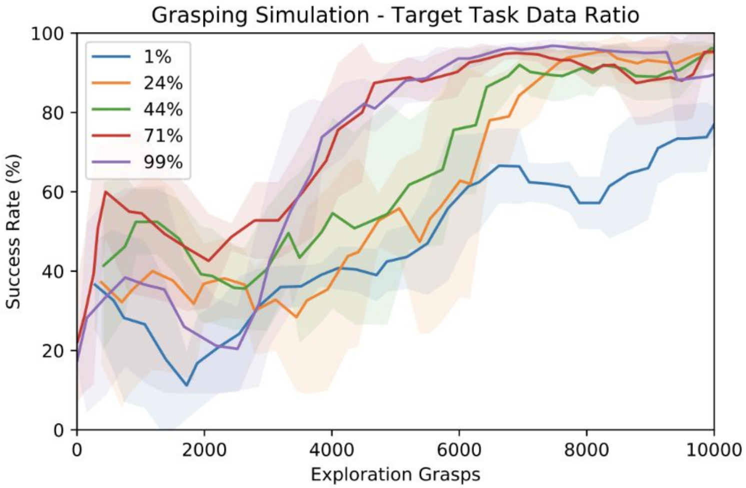


Source: An Overview of Multi-Task Learning in Deep Neural Networks. Ruder, 2017.

Fine-Tuning for Off-Policy RL (vs. Supervised)

Techniques Studied (What didn't work)

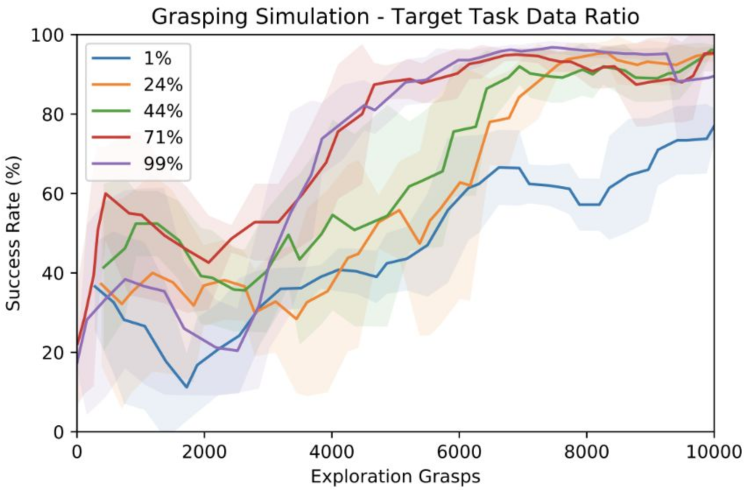
- Architectural
  - Adding a Q-function head
  - Training only some layers (front, middle, back, etc.)
  - Re-initializing some layers
  - Training only batch norms
  - etc.
- Sampling
  - Different sampling probability of old/new data
  - Using n-step returns (to get supervision info out of same data)
- **What was important**
  - Gradients per new sample
  - Learning rate



Fine-Tuning for Off-Policy RL (vs. Supervised)

What does work

- Continue training the entire network
- (there is no second bullet)

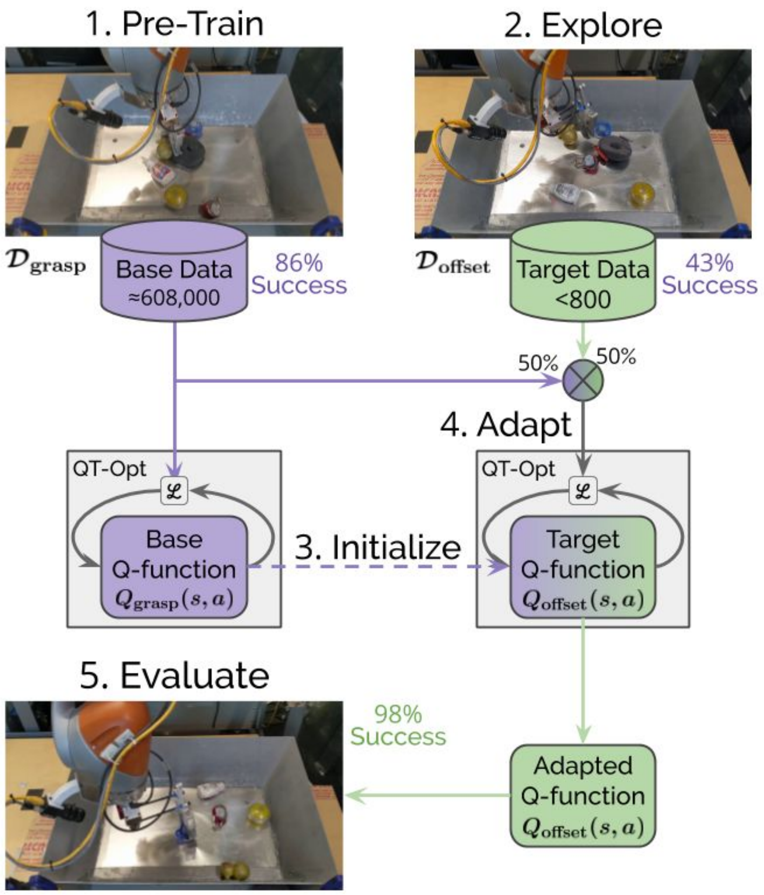


Roadmap

- Problem
- Preliminaries
- Baseline Study
- Fine-Tuning for Off-Policy RL
- **A Very Simple Fine-Tuning Method**
- From Fine-Tuning to Continual Learning
- Insights and Issues

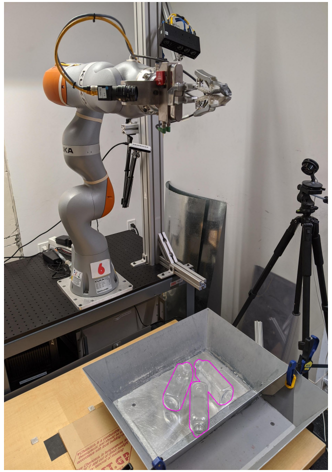
A Very Simple Method

- Fine-tuning method
  - **Pre-Train:** Pre-trained policy, pre-training data
  - **Explore** using the pre-trained policy (e.g. vanilla grasping)
  - **Initialize** QT-Opt with pre-trained policy (Q-function), pre-training data, new data
  - **Adapt** pre-trained policy using RL select new vs. old data with 50% probability
  - **Evaluate** updated policy on robot
- **Completely offline**



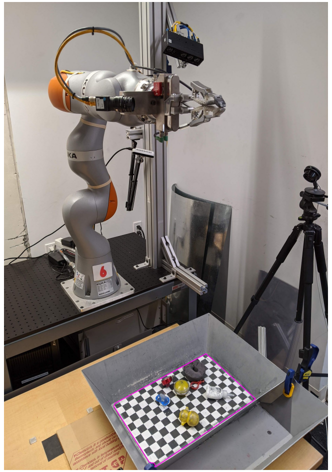
A Very Simple Method: Experiments

49%



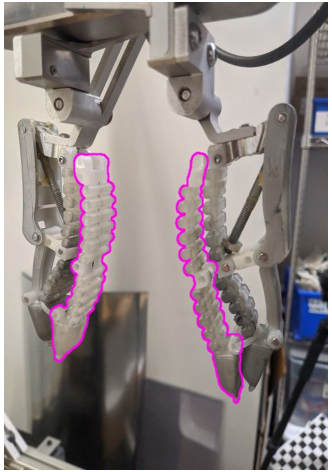
Transparent Bottles

50%



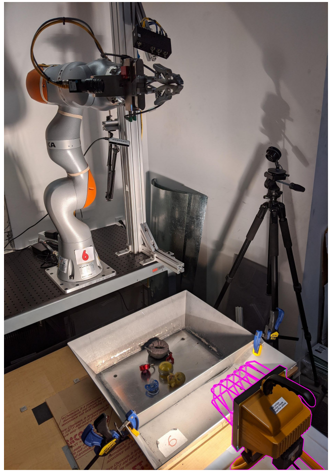
Checkerboard Backing

75%



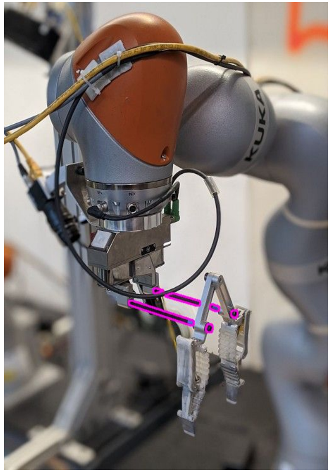
Extend Gripper 1cm

32%



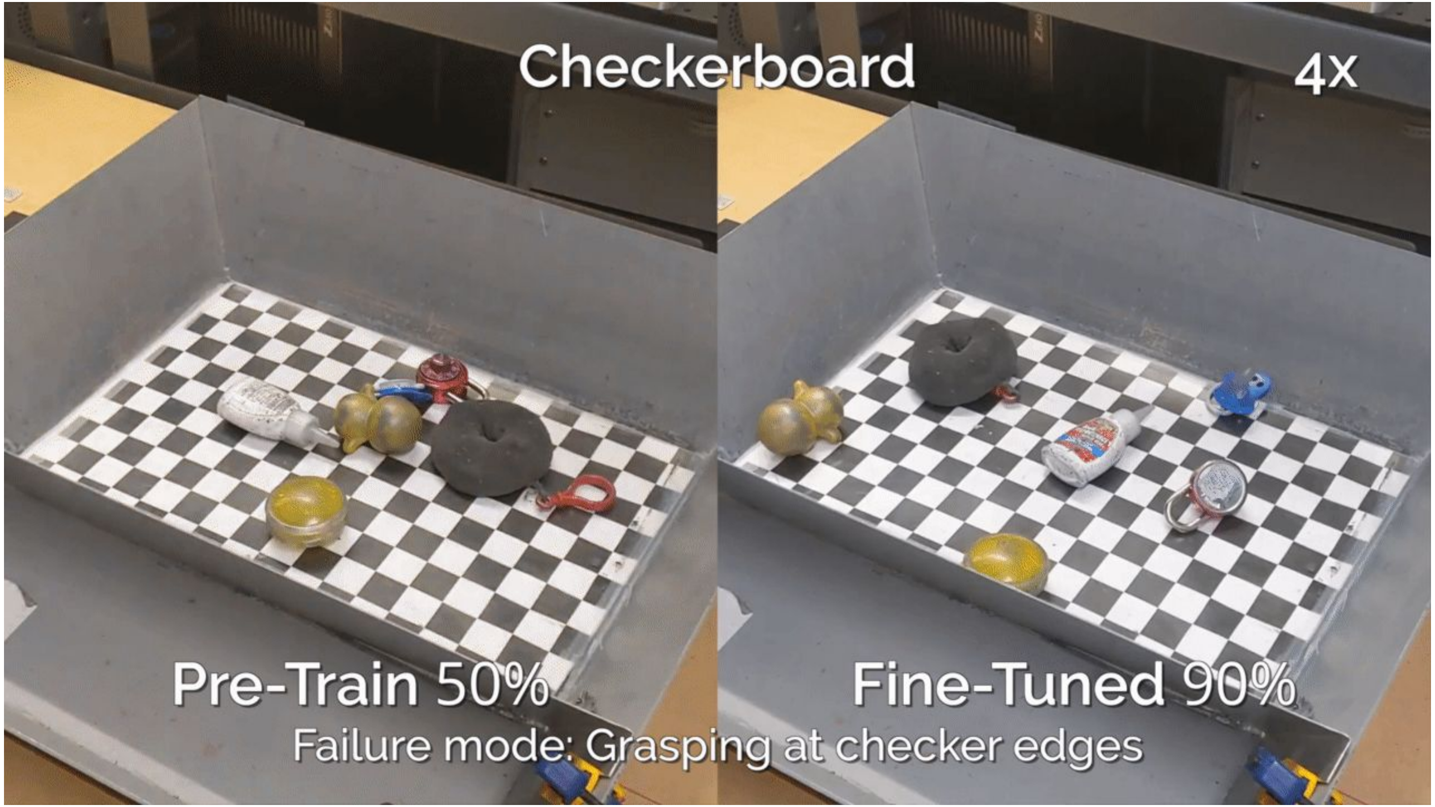
Harsh Lighting

43%

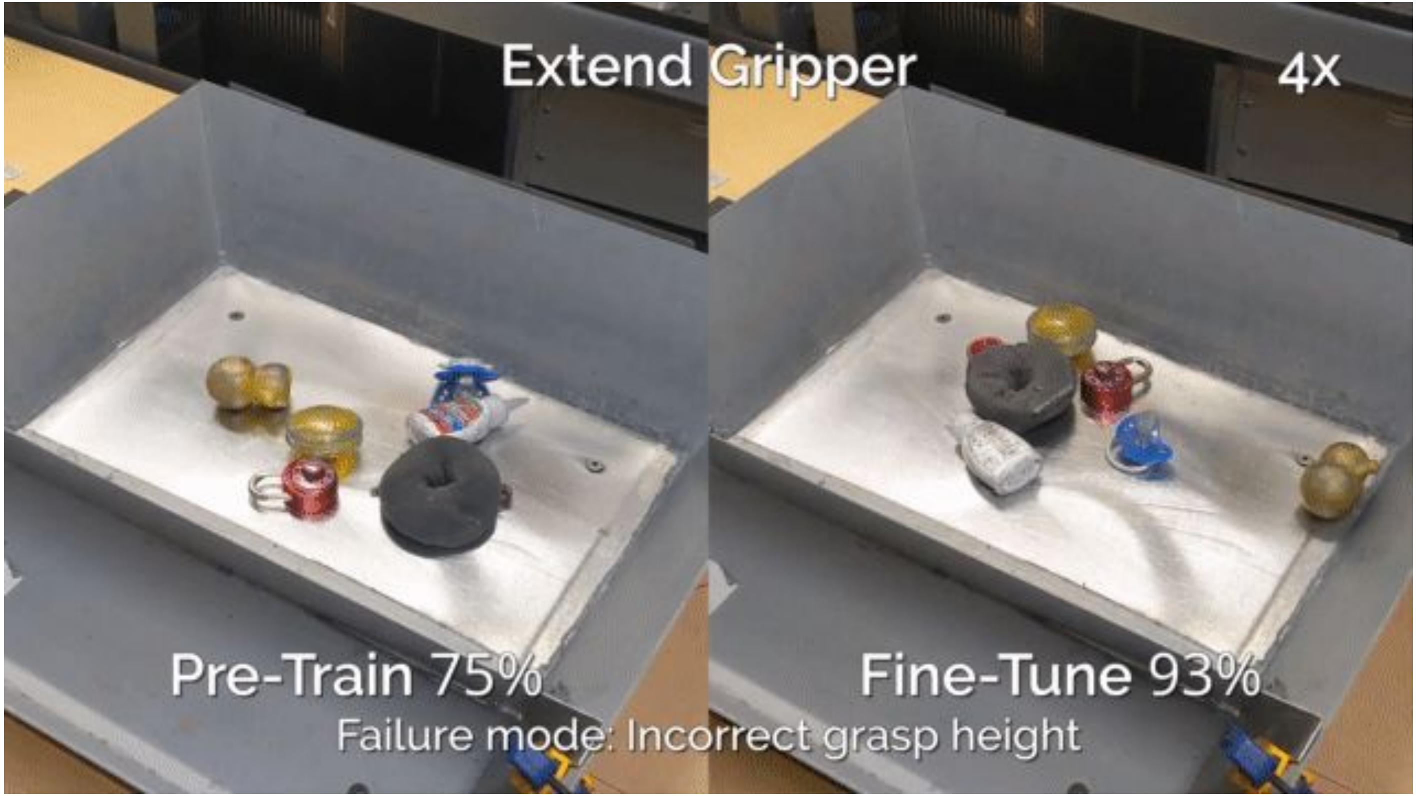


Offset Gripper 10cm

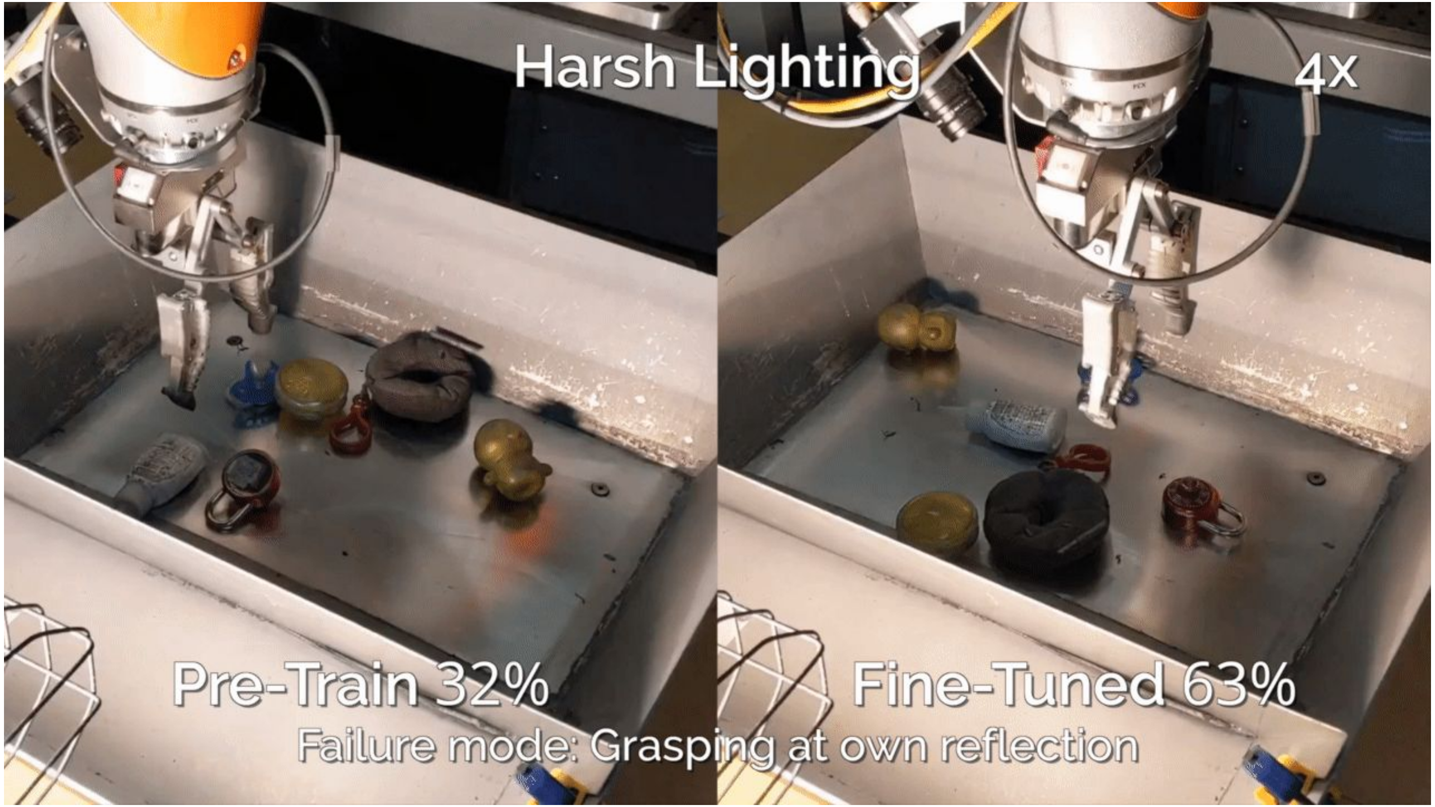
A Very Simple Method: Results



A Very Simple Method: Results

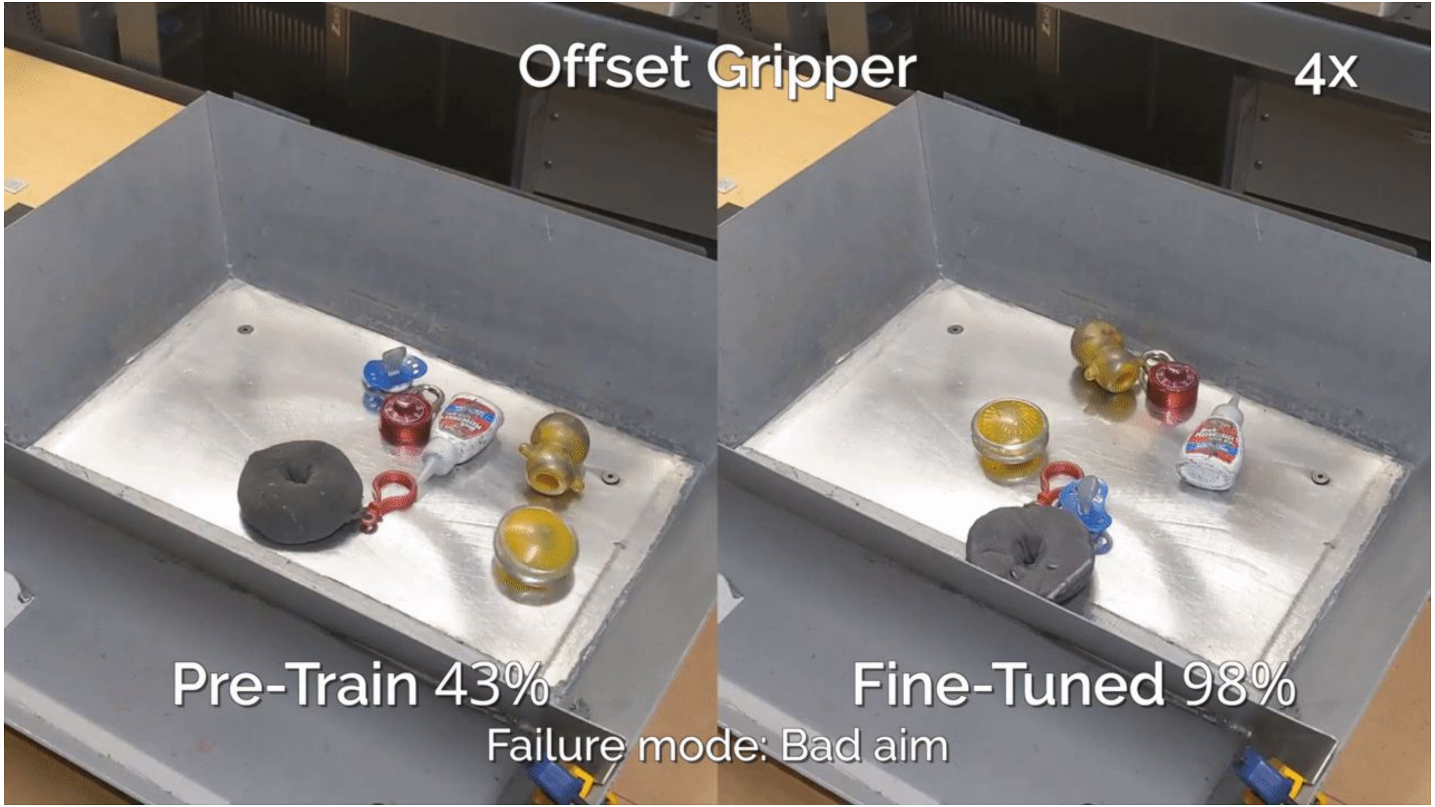


A Very Simple Method: Results

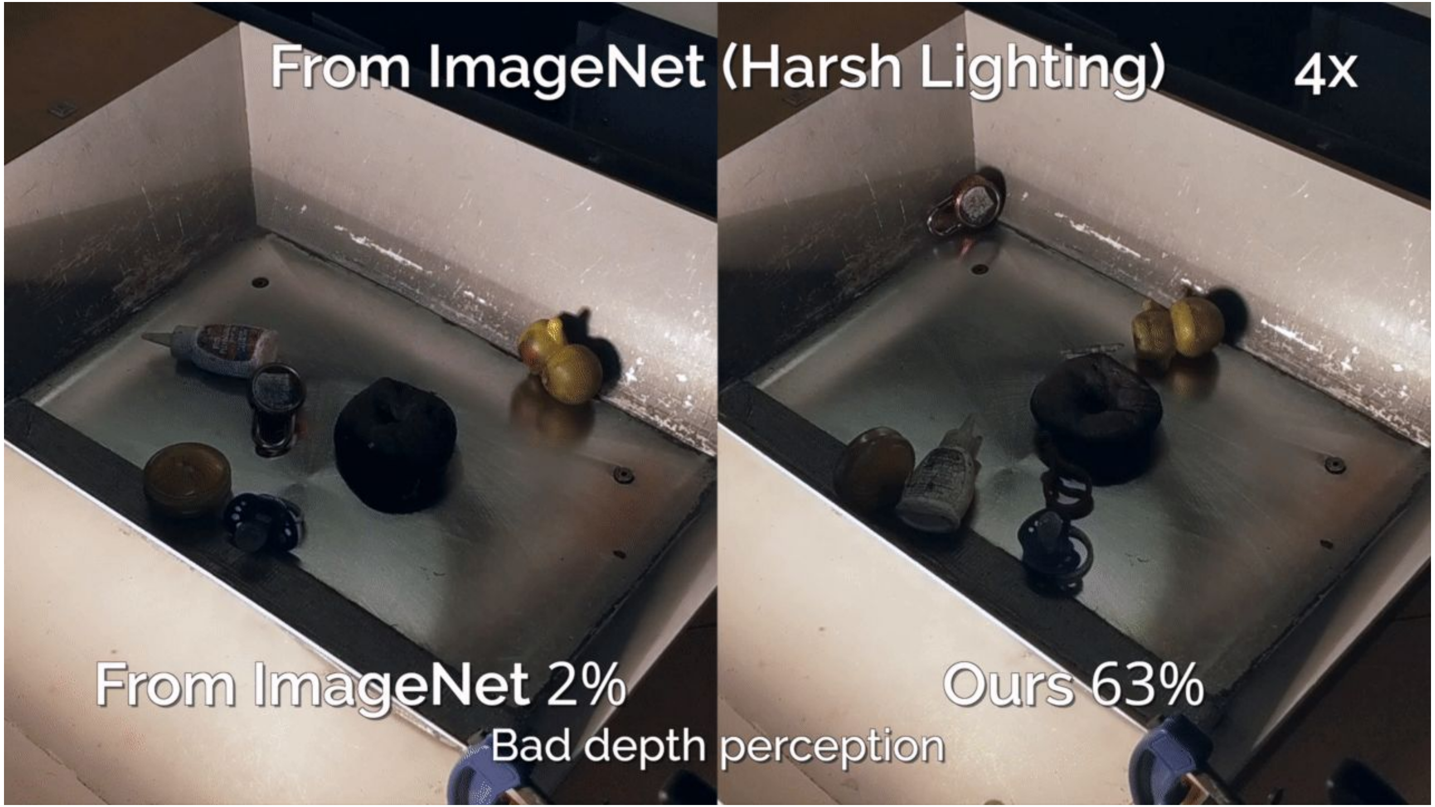


A Very Simple Method: Results

A Very Simple Method: Results

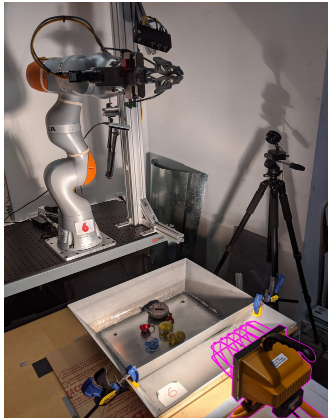


A Very Simple Method: RL Matters



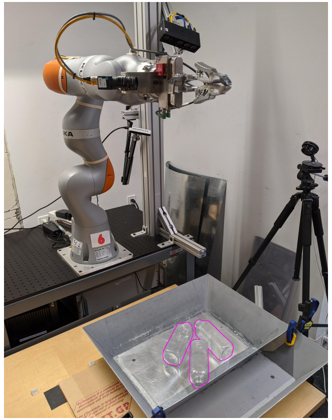
A Very Simple Method: Results

32% → **63%**



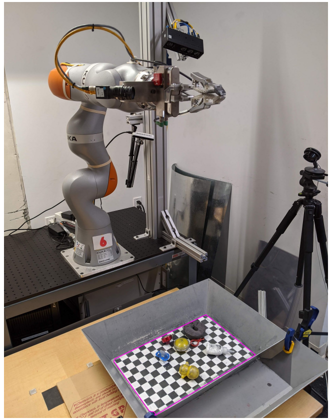
Harsh Lighting

49% → **66%**



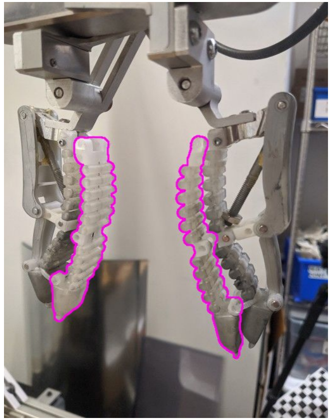
Transparent Bottles

50% → **90%**



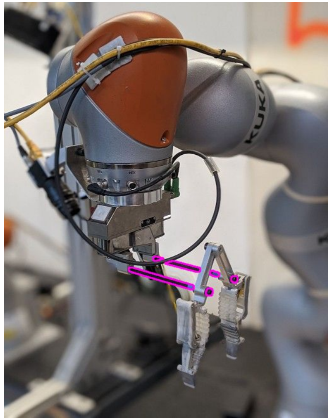
Checkerboard Backing

75% → **93%**



Extend Gripper 1cm

43% → **98%**



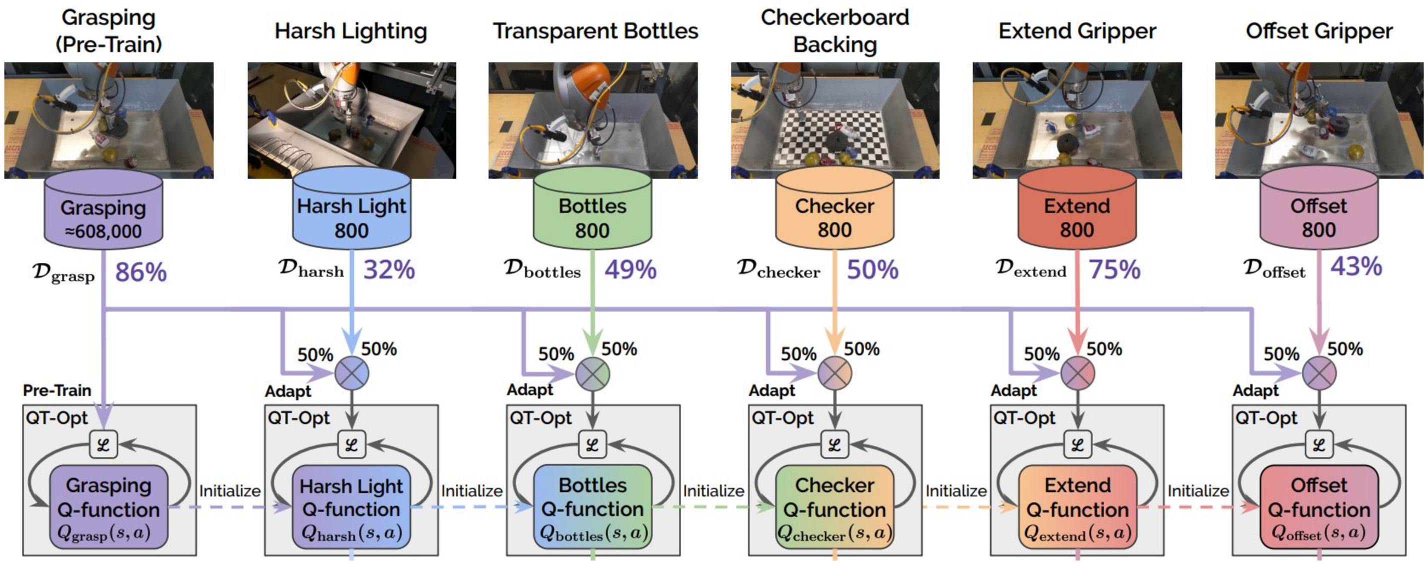
Offset Gripper 10cm

Roadmap

- Problem
- Preliminaries
- Baseline Study
- Fine-Tuning for Off-Policy RL
- A Very Simple Fine-Tuning Method
- **From Fine-Tuning to Continual Learning**
- Insights and Issues

Continual Learning: Experiment

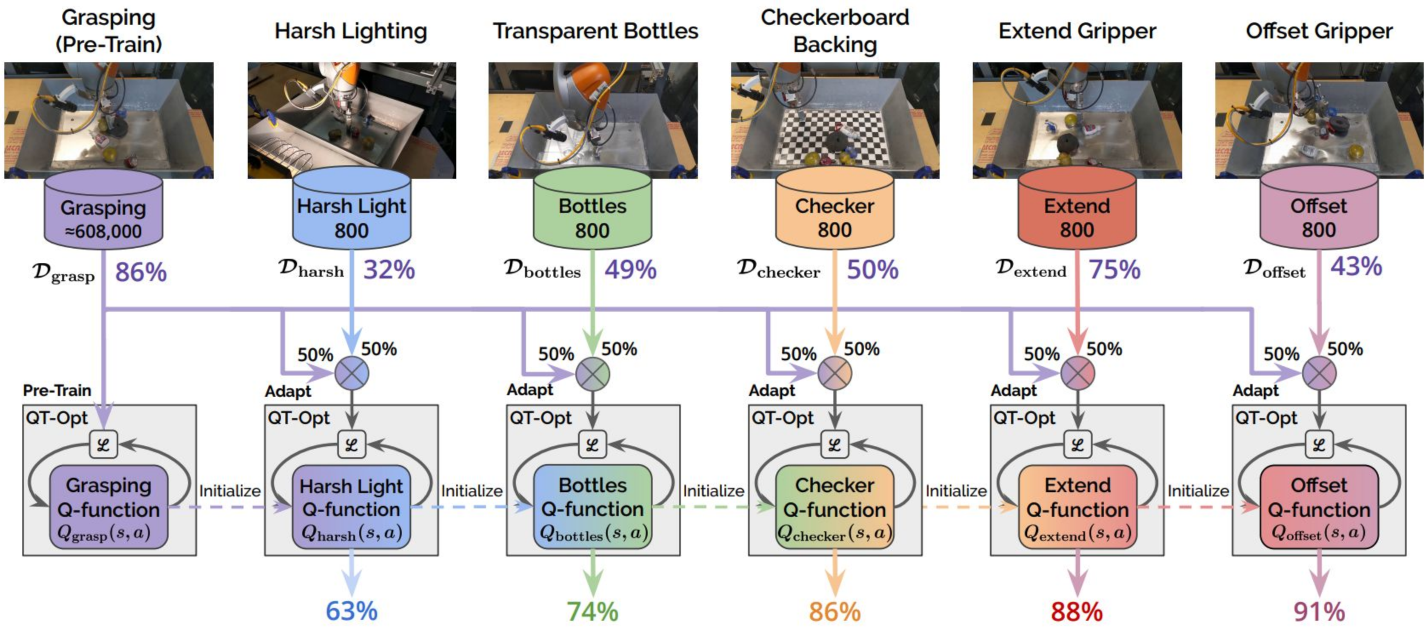
Re-train a single lineage of policies repeatedly



Continual Learning: Results

Continual Learning: Results

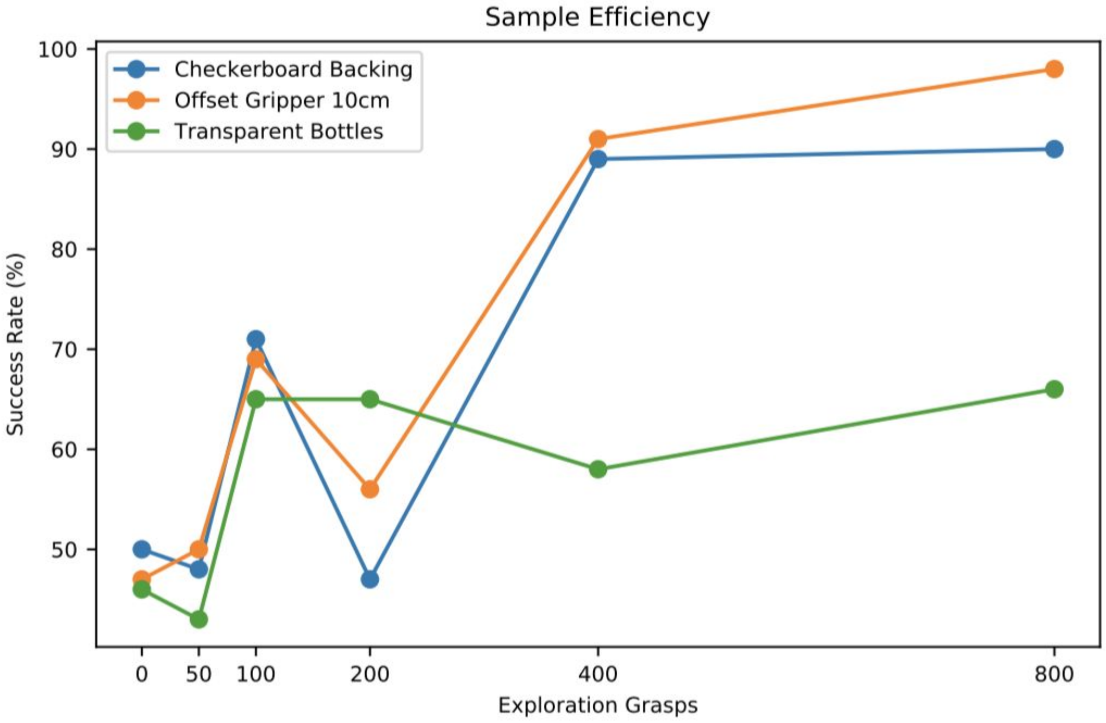
Continual Learning: Results



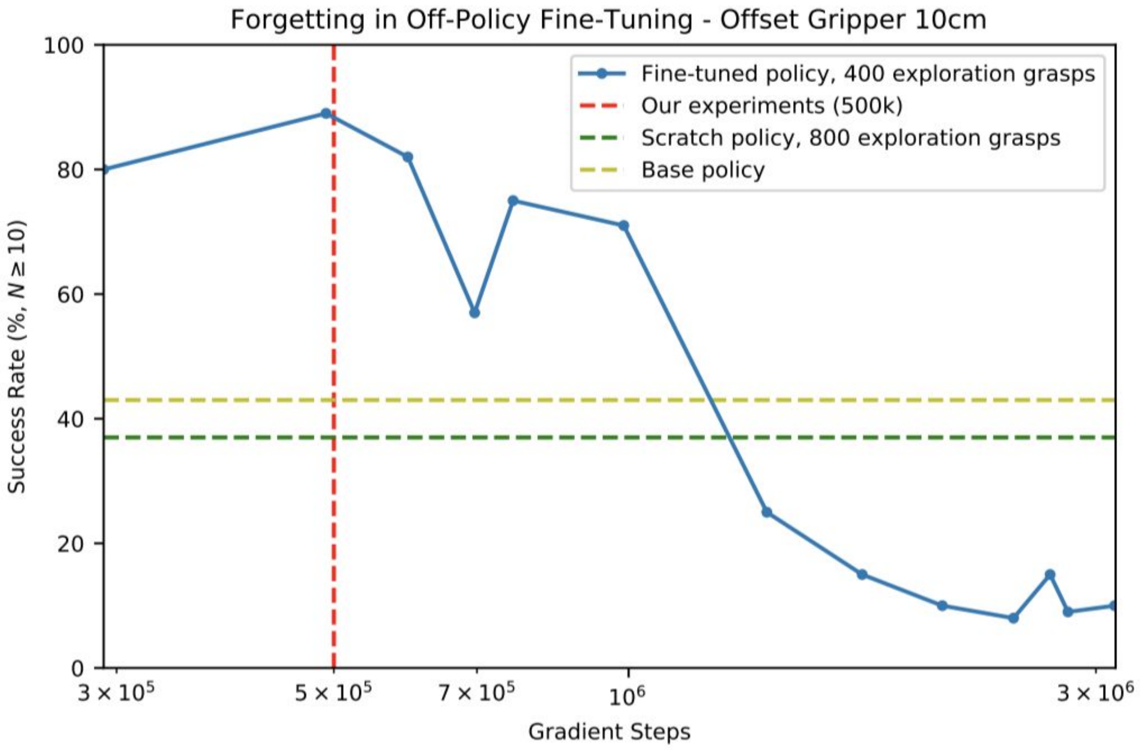
Roadmap

- Problem
- Preliminaries
- Baseline Study
- Fine-Tuning for Off-Policy RL
- A Very Simple Fine-Tuning Method
- From Fine-Tuning to Continual Learning
- **Insights and Issues**

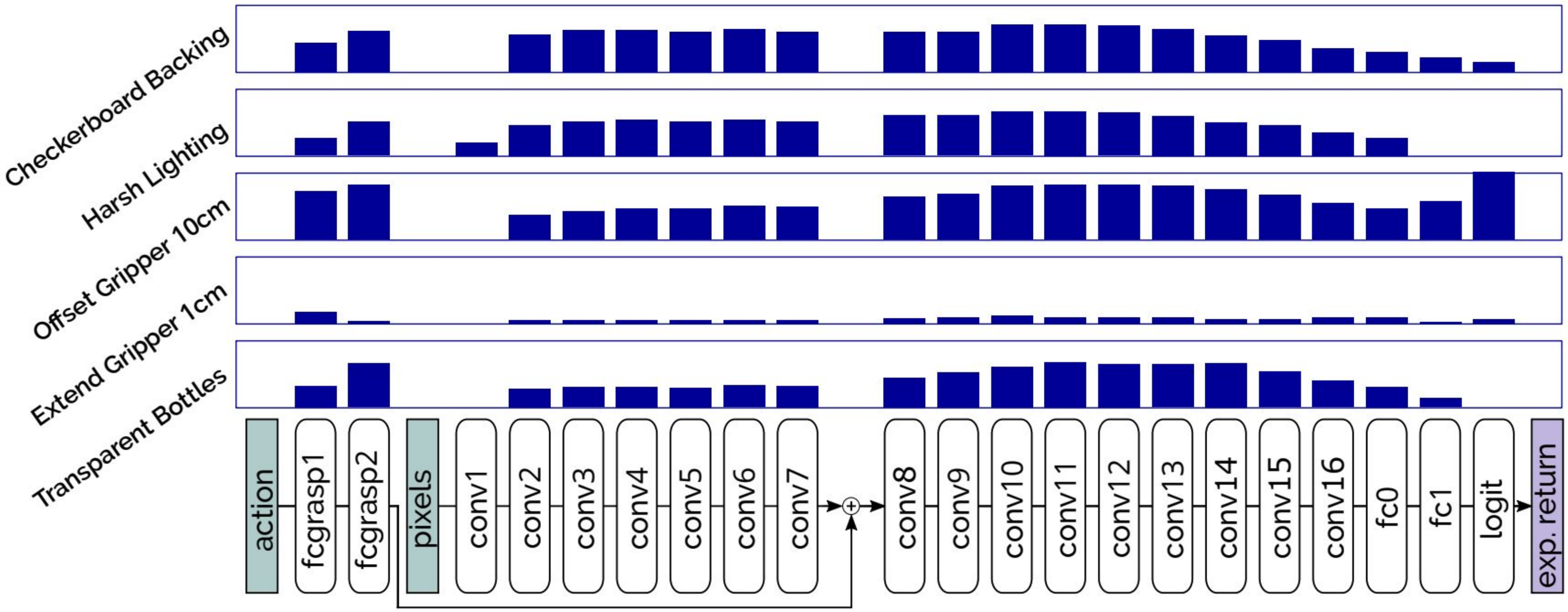
Insights and Issues: Sample Efficiency



Insights and Issues: Knowing when to stop



Insights and Issues: What gets updated?



Conclusions

Offline fine-tuning: A promising building block for continual learning

- **Fast**
  1-4 hours of practice, 0.2%
- **Simple**
  Barely different from regular training
- **Repeatable**
  Works in a continual setting with ~0% performance penalty

Future Directions

- How extreme are the target tasks can we adapt to?
  → off-distribution and structural adaptation
- Can we choose to explore (vs. exploit) automatically?
  → off-policy evaluation
- Can we integrate this to create a fully automatic learner?
  → lifelong and continual learning

Thank You!

- Collaborators: Karol Hausman, Chelsea Finn, Sergey Levine, Ben Swanson
- Adviser: Gaurav Sukhatme
- CoRL organizers and reviewers

More Info

- Visit the website: https://ryanjulian.me/never-stop-learning
- Read the paper: https://arxiv.org/abs/2004.10190
- Watch the video: https://youtu.be/pPDVewcSpdc
- Contact me: ryanjulian@gmail.com / https://ryanjulian.me

Challenge Task	Original Policy	Ours (exploration grasps)							Comparisons	
		25	50	100	200	400	800	**Best** (Δ)	Scratch	ImageNet
Checkerboard Backing	50%	67%	48%	71%	47%	89%	90%	**90% (+40)**	0%	0%
Harsh Lighting	32%	23%	16%	52%	44%	58%	63%	**63% (+31)**	4%	2%
Extend Gripper 1 cm	75%	93%	67%	80%	51%	90%	69%	**93% (+18)**	0%	14%
Offset Gripper 10 cm	43%	73%	50%	60%	56%	91%	98%	**98% (+55)**	37%	47%
Transparent Bottles	49%	46%	43%	65%	65%	58%	66%	**66% (+17)**	27%	20%
Baseline Grasping Task	86%	98%	81%	84%	78%	93%	89%	**98% (+12)**	0%	12%

← Every cell is a ~1 hr experiment!

Questions?