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
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

Preliminaries: QT-Opt

- replay buffers
  - offline data: 580K grasps
  - off-policy
  - on-policy
  - train
- Bellman Updater $Q_T$
- Training Worker $E_q(1)$
- Cross Entropy Method
  \[
  \pi(s) = \arg \max_a Q_\theta(s, a)
  \]

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
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
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
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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transparent Bottles</td>
<td>49%</td>
</tr>
<tr>
<td>Checkerboard Backing</td>
<td>50%</td>
</tr>
<tr>
<td>Extend Gripper 1cm</td>
<td>75%</td>
</tr>
<tr>
<td>Harsh Lighting</td>
<td>32%</td>
</tr>
<tr>
<td>Offset Gripper 10cm</td>
<td>43%</td>
</tr>
</tbody>
</table>
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

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

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

- Transparent Bottles: 49%
- Checkerboard Backing: 50%
- Extend Gripper 1cm: 75%
- Harsh Lighting: 32%
- Offset Gripper 10cm: 43%
A Very Simple Method: Results

Checkerboard 4X

Pre-Train 50%
Failure mode: Grasping at checker edges

Fine-Tuned 90%
A Very Simple Method: Results

Pre-Train 75%
Failure mode: Incorrect grasp height

Fine-Tune 93%

Extend Gripper 4X
A Very Simple Method: Results

Harsh Lighting
Pre-Train 32%
Failure mode: Grasping at own reflection

Fine-Tuned 63%
A Very Simple Method: Results
A Very Simple Method: Results

Offset Gripper

4x

Pre-Train 43%
Failure mode: Bad aim

Fine-Tuned 98%
A Very Simple Method: RL Matters
A Very Simple Method: Results

32% → 63%
49% → 66%
50% → 90%
75% → 93%
43% → 98%

Harsh Lighting
Transparent Bottles
Checkerboard Backing
Extend Gripper 1cm
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
- Watch the video: https://youtu.be/pPDVewcSpdc
- Contact me: ryanjulian@gmail.com / https://ryanjulian.me

<table>
<thead>
<tr>
<th>Challenge Task</th>
<th>Original Policy</th>
<th>Ours (exploration grasps)</th>
<th>Best (Δ)</th>
<th>Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Checkerboard Backing</td>
<td>50%</td>
<td>67%</td>
<td>48%</td>
<td>71%</td>
</tr>
<tr>
<td>Harsh Lighting</td>
<td>32%</td>
<td>23%</td>
<td>16%</td>
<td>52%</td>
</tr>
<tr>
<td>Extend Gripper 1 cm</td>
<td>75%</td>
<td>93%</td>
<td>67%</td>
<td>80%</td>
</tr>
<tr>
<td>Offset Gripper 10 cm</td>
<td>43%</td>
<td>73%</td>
<td>50%</td>
<td>60%</td>
</tr>
<tr>
<td>Transparent Bottles</td>
<td>49%</td>
<td>46%</td>
<td>43%</td>
<td>65%</td>
</tr>
<tr>
<td>Baseline Grasping Task</td>
<td>86%</td>
<td>98%</td>
<td>81%</td>
<td>84%</td>
</tr>
</tbody>
</table>

← Every cell is a ~1 hr experiment!
Questions?