One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning

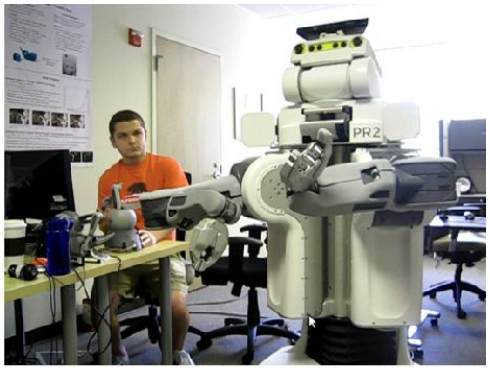
Authors: Tianhe Yu*, Chelsea Finn*, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, Sergey Levine

CS 330: Deep Multi-Task and Meta-Learning
October 16, 2019

Problem: Imitation learning

Problem: Imitation learning

(a) Kinesthetic Teaching

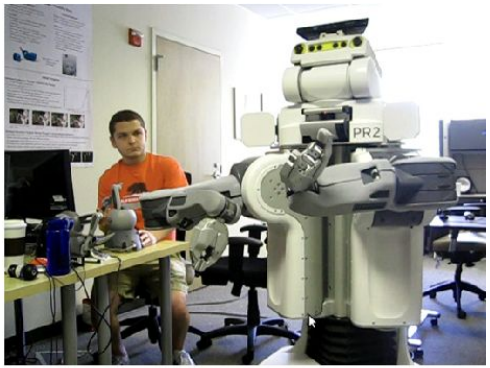


(b) Teleoperation

Learning from direct manipulation of actions

Problem: Imitation learning

(a) Kinesthetic Teaching

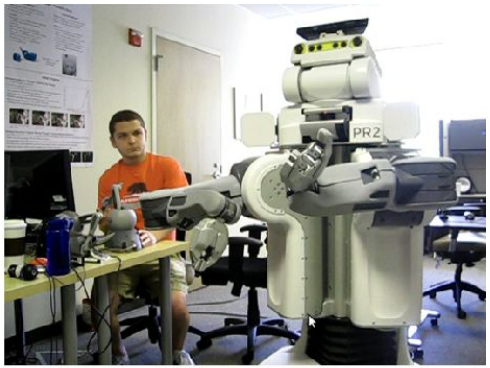


(b) Teleoperation

Learning from direct manipulation of actions vs Learning from visual input

Problem: Imitation learning

(a) Kinesthetic Teaching



(b) Teleoperation

1) Visual-based imitation learning often **requires a large number of human demonstrations**

2) Must deal with **domain-shift** between different demonstrators, objects, backgrounds, as well as correspondence between human and robot body parts

Goal

Meta-learn a prior such that…

- The robot can learn to **manipulate new objects** after seeing a **single** video of a human demonstration (*one-shot imitation*)

- The robot can **generalize** to human demonstrations from different backgrounds and morphologies (*domain shift*)

Domain-Adaptive Meta-Learning

End goal

One human demonstration

Infer robot policy from human demo

Domain-Adaptive Meta-Learning

Human demonstrations

$\mathbf{d}^h = \mathbf{o}_1, ..., \mathbf{o}_T$

Robot demonstrations

$\mathbf{d}^r = \mathbf{o}_1, \mathbf{s}_1, \mathbf{a}_1, ..., \mathbf{o}_T, \mathbf{s}_T, \mathbf{a}_T$

Learn ***how*** to infer robot policy from human demo

End goal

One human demonstration

Infer robot policy from human demo

Domain-Adaptive Meta-Learning

Human demonstrations

$\mathbf{d}^h = \mathbf{o}_1, ..., \mathbf{o}_T$

Robot demonstrations

$\mathbf{d}^r = \mathbf{o}_1, \mathbf{s}_1, \mathbf{a}_1, ..., \mathbf{o}_T, \mathbf{s}_T, \mathbf{a}_T$

Learn ***how*** to infer robot policy from human demo

End goal

One human demonstration

Infer robot policy from human demo

Domain-Adaptive Meta-Learning

Human demonstrations

$\mathbf{d}^h = \mathbf{o}_1, ..., \mathbf{o}_T$

Robot demonstrations

$\mathbf{d}^r = \mathbf{o}_1, \mathbf{s}_1, \mathbf{a}_1, ..., \mathbf{o}_T, \mathbf{s}_T, \mathbf{a}_T$

Learn ***how*** to infer robot policy from human demo

$$\min_{\theta,\psi} \sum_{\mathcal{T}\sim p(\mathcal{T})} \sum_{\mathbf{d}^h \in \mathcal{D}^h_{\mathcal{T}}} \sum_{\mathbf{d}^r \in \mathcal{D}^r_{\mathcal{T}}} \mathcal{L}_{\text{BC}}(\theta - \alpha\nabla_\theta \mathcal{L}_\psi(\theta, \mathbf{d}^h), \mathbf{d}^r)$$

End goal

One human demonstration

Infer robot policy from human demo

$\phi_{\mathcal{T}} = \theta - \alpha\nabla_\theta \mathcal{L}_\psi(\theta, \mathbf{d}^h)$

Problem Definition and Terminology

Goal is to infer the robot policy parameters $\phi_{\mathcal{T}_i}$ that will accomplish the task $\mathcal{T}_i$

- Learn prior that encapsulates visual and physical understanding of the world using **human** and **robot** demonstration data from a variety of tasks
- $\mathbf{d}^h = \mathbf{o}_1, \ldots, \mathbf{o}_T$ (sequence of human observations)
- $\mathbf{d}^r = \mathbf{o}_1, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{o}_T, \mathbf{s}_T, \mathbf{a}_T$ (sequence of robot observations, states, and actions)

Meta-training algorithm

Input: *Human and robot demos for tasks from* $p(\mathcal{T})$

while training **do**:

1. Sample task
2. Sample **human demo**
3. Compute *policy* params
4. Sample **robot demo**
5. Update *meta* params

$\mathcal{T} \sim p(\mathcal{T})$

$\mathbf{d}^h \sim \mathcal{D}^h_{\mathcal{T}}$ **INNER LOOP**

$$\phi_{\mathcal{T}} = \theta - \alpha \nabla_\theta \mathcal{L}_\psi(\theta, \mathbf{d}^h)$$

$\mathbf{d}^r \sim \mathcal{D}^r_{\mathcal{T}}$ **OUTER LOOP**

$$(\theta, \psi) \leftarrow (\theta, \psi) - \beta \nabla_{\theta,\psi} \mathcal{L}_{\mathrm{BC}}(\phi_{\mathcal{T}}, \mathbf{d}^r)$$

Output: *Meta params* θ, ψ

(HOW to infer robot policy from human demo)

Meta-test algorithm

Input:

1. *Meta-learned initial policy params* θ
2. *Meta-learned adaptation objective* $\mathcal{L}_\psi$
3. *One video of* ***human demo*** *for a new task* $\mathbf{d}^h \sim \mathcal{D}_\mathcal{T}^h$

Compute policy params via one gradient step

$$\phi_\mathcal{T} = \theta - \alpha \nabla_\theta \mathcal{L}_\psi(\theta, \mathbf{d}^h)$$

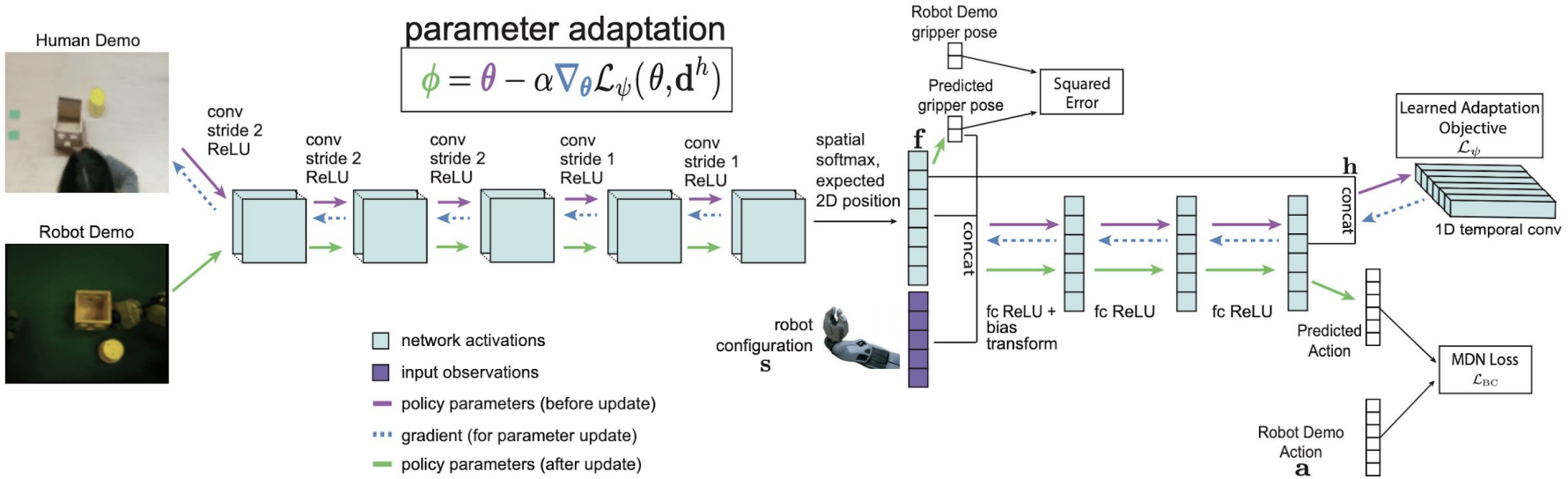
INNER LOOP

Output: *Policy params*

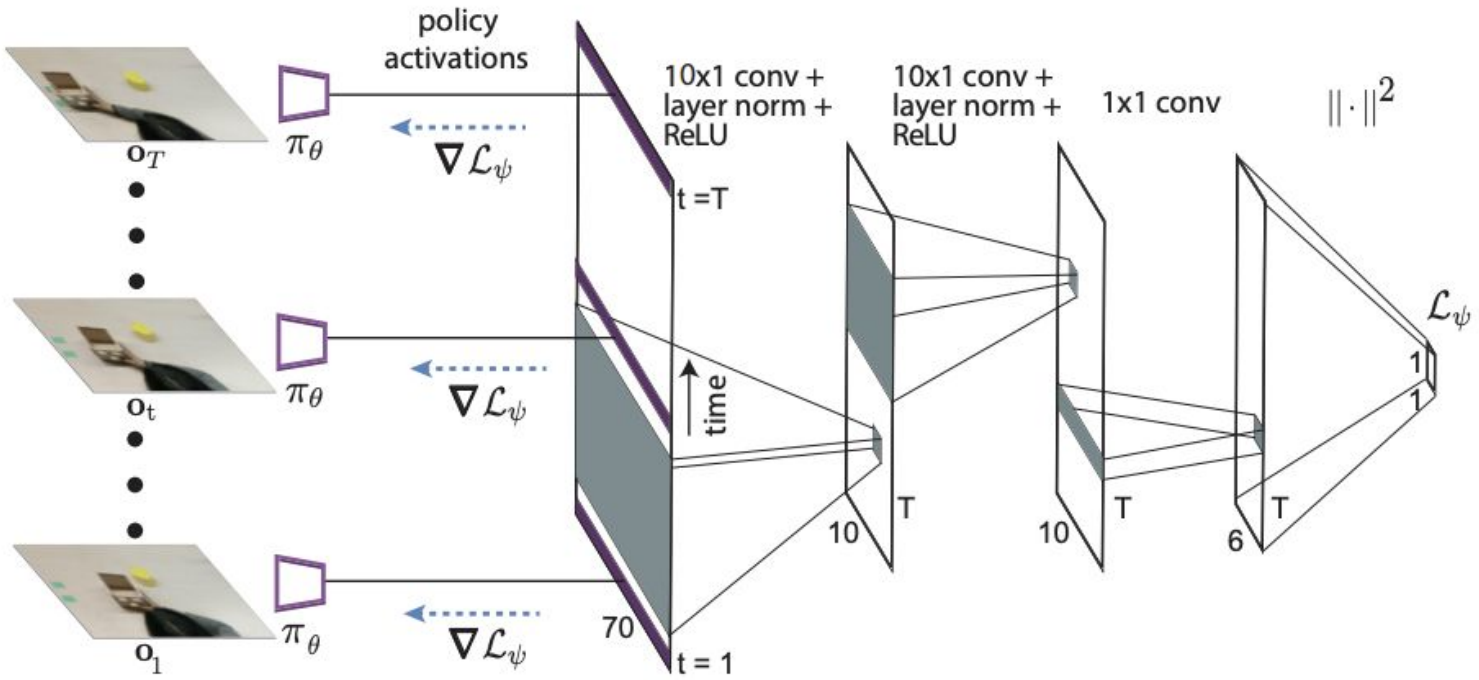
(robot policy inferred from human demo)

Architecture Overview

$$\min_{\theta,\psi} \sum_{\mathcal{T}\sim p(\mathcal{T})} \sum_{\mathbf{d}^h \in \mathcal{D}^h_{\mathcal{T}}} \sum_{\mathbf{d}^r \in \mathcal{D}^r_{\mathcal{T}}} \mathcal{L}_{\text{BC}}(\theta - \alpha \nabla_\theta \mathcal{L}_\psi(\theta, \mathbf{d}^h), \mathbf{d}^r)$$



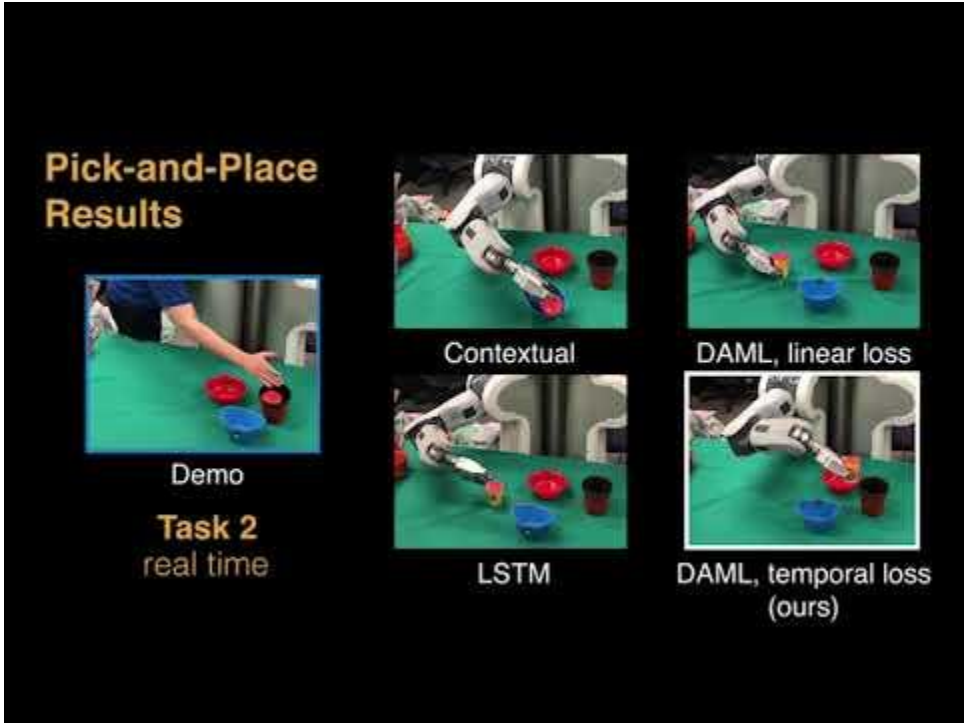
Learned temporal adaptation objective $\mathcal{L}_\psi$



Compared meta-learning approaches:

- Contextual policy
- DA-LSTM policy (Duan. et al.)
- DAML (linear loss)
- **DAML (temporal loss)**

Results (video)



Exp. 1) Placing, Pushing, and Pick & Place using PR2

- Using human demonstrations from the perspective of the robot

	placing	pushing	pick and place
DA-LSTM	33.3%	33.3%	5.6%
contextual	36.1%	16.7%	16.7%
DAML, linear loss	76.7%	27.8%	11.1%
DAML, temporal loss (ours)	**93.8%**	**88.9%**	**80.0%**

Exp. 2) Pushing Task with Large Domain Shift using PR2

- Using human demonstrations collected in a different room with a different camera and camera perspective from that of the robot

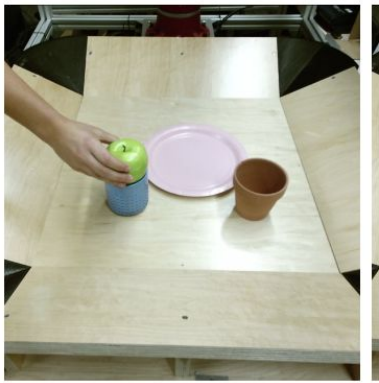
Critique: Does not explore capability of handling domain shift of baselines

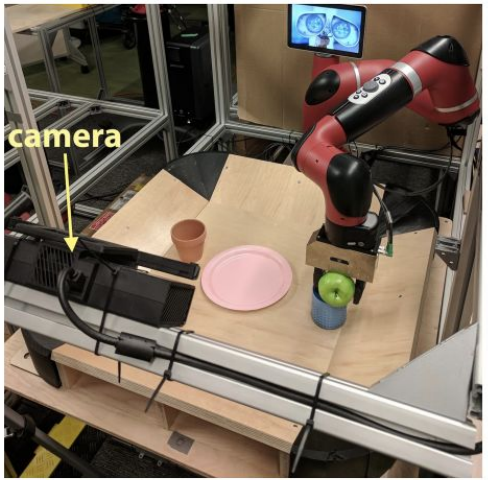
pushing	seen bg	novel bg 1	novel bg 2
DAML, temporal loss (ours)	**81.8%**	**66.7%**	**72.7%**

Failure analysis of DAML	seen bg	novel bg 1	novel bg 2
# successes	27	22	24
# failures from task identification	1	5	4
# failures from control	5	6	5

Exp. 3) Placing using Sawyer

- Used kinesthetic teaching instead of teleoperation for outer loss
- Assessing generality on a different robot and a different form of robot demonstration collection
- 77.8% placing success rate





Exp. 4) Learned Adaptation Objective on Pushing Task

- Experiment performed in simulation and without domain shift to isolate temporal adaptation loss

	simulated pushing no domain shift
LSTM [10]	34.23%
contextual	56.98%
MIL, linear loss [15]	66.44%
MIL, temporal loss (ours)	**80.63%**

Strengths + Takeaways

- Success on **one-shot imitation** from **visual** input of human demonstrations
  - Extension of **MAML** to **domain adaptation** by defining inner loss using **policy activations** rather than actions
  - Learned **temporal adaptation** objective that exploits temporal information
  - Can do this even if human demonstration video is from a substantially different setting
- Performs well even though **amount of data per task is low**
  - Can adapt to a **diverse range of tasks**

Limitations

- Has not demonstrated ability to learn ***entirely new*** motions
  - Domain-shift due to new background, demonstrator, viewpoint etc. was handled, but the actual behaviors at meta-test time are structurally similar to those at meta-training time
- More data during meta-training could enable better results
  - Few thousand demonstrations but total amount of data per task is quite low
- Still **requires robot demos** (paired up with human demos)
  - Has not yet solved the problem of learning purely from human demos without the need for any training using robot demos

Discussion questions

- How should we interpret the meta-learned temporal adaptation objective $\mathcal{L}_\psi$?
  - What does this meta-learned loss represent? How can we make it more interpretable?
- Can this approach be extended to tasks with more complex actions?
  - Is meta-learning a loss on policy activations instead of explicitly computing the loss on actions sufficient for more complex tasks?