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

Presentation by Siavash Khodadadeh
Outline

- Meta Learning
- Problem Definition
- Network Architecture
- Experiments
- Conclusions
One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning

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

*denotes equal contribution
One-Shot Imitation from Observing Humans

- Learn from very few demonstrations
- Instantly generalize to new situations

Human Demo

Robot Demo
Approaches for Imitation Learning

- Supervised Learning
  - Backpropagation
  - Lots of demonstrations.
  - Different Objects?
Approaches for Imitation Learning

- **Time Contrastive Networks**
  - Learn Embeddings
  - Lots of demonstrations.
  - Different objects?
Approaches for Imitation Learning

- Meta Learning
  - Learn to learn
    - Using prior knowledge built up from many other tasks
    - Learn the new task from just one demonstration
Dataset

Task 1
Push left the pack

Task 2
Push left the cup

One-Shot Learning

Push left the bottle
Problem Definition

Set of tasks \( \{ i \} \) \( \{ (D^h_{\tau_i}, D^r_{\tau_i}) \} \)

- \( d^h \): \((o_1, o_2, \ldots, o_T)\)
- \( d^r \): \((o_1, s_1, a_1, \ldots, o_T, s_T, a_T)\)
  - States: Joint angles
  - Actions:
    - Torque applied to joints
    - Gripper
Problem Definition

- Policy $\pi_{\theta}$
  - $s_i, o_i \rightarrow$ action
  - (parameterized by $\theta$)

- Loss $\mathcal{L}_{BC}$ over policy for demonstration $d^r$

\[
\mathcal{L}_{BC}(\theta, d^r) = \mathcal{L}_{BC}(\theta, \{o_{1:T}, s_{1:T}, a_{1:T}\}) = \sum_t \log \pi_{\theta}(a_t | o_t, s_t)
\]
Problem Definition

\[ \mathcal{L}_{BC}(\theta, d^r) = \mathcal{L}_{BC}(\theta, \{o_{1:T}, s_{1:T}, a_{1:T}\}) = \sum_t \log \pi_\theta(a_t | o_t, s_t) \]

\[ \mathcal{L}_\psi(\theta, d^h) = \text{Policy loss for human demonstration} \]

We will learn it through meta-training
Training Objective

- Different tasks
  - For each task $\tau$:
    - Human demonstrations $d^h$
    - Robot demonstrations $d^r$

$$\min_{\theta,\psi} \sum_{\tau \sim p(\tau)} \sum_{d^h \in D^h_\tau} \sum_{d^r \in D^r_\tau} L_{BC}(\theta - \alpha \nabla_{\theta} L_{\psi}(\theta, d^h), d^r)$$
The Policy Network
Algorithm 1 Meta-imitation learning from humans

Require: $\{(D^h_{\mathcal{T}_i}, D^r_{\mathcal{T}_i})\}$: human and robot demonstration data for a set of tasks $\{\mathcal{T}_i\}$ drawn from $p(\mathcal{T})$

Require: $\alpha, \beta$: inner and outer step size hyperparameters

while training do
  Sample task $\mathcal{T} \sim p(\mathcal{T})$ \{or minibatch of tasks\}
  Sample video of human $d^h \sim D^h_{\mathcal{T}}$
  Compute policy parameters $\phi_{\mathcal{T}} = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\psi}(\theta, d^h)$
  Sample robot demo $d^r \sim D^r_{\mathcal{T}}$
  Update $(\theta, \psi) \leftarrow (\theta, \psi) - \beta \nabla_{\theta, \psi} \mathcal{L}_{BC}(\phi_{\mathcal{T}}, d^r)$
end while

Return $\theta, \psi$
Algorithm 2 Learning from human video after meta-learning

Require: meta-learned initial policy parameters $\theta$
Require: learned adaptation objective $L_\psi$
Require: one video of human demo $d^h$ for new task $\mathcal{T}$

Compute policy parameters $\phi_\mathcal{T} = \theta - \alpha \nabla_\theta L_\psi(\theta, d^h)$
return $\pi_\phi$
The Policy Network

Parameter Adaptation:
\[ \phi = \theta - \alpha \nabla_{\theta} \mathcal{L}_\psi(\theta, d^h) \]

Learned Adaptation Objective \( \mathcal{L}_\psi \)

Squared Error

1D Temporal Conv

Extracts 2D Points

Concat

50 fc ReLU
Bias transform

Predicted Action \( a \)

Robot Demo

Action

Robot Demo gripper pose

Predicted gripper pose

 concatenation

Robot Demo gripper pose

concatenation

network activations
input observations
policy parameters (before update)
gradient (for parameter update)
policy parameters (after update)
Temporal Convolution Network

\[
\begin{align*}
    &h_1 \quad \nabla \mathcal{L}_\psi \\
    &h_t \quad \nabla \mathcal{L}_\psi \\
    &h_T \quad \nabla \mathcal{L}_\psi \\
\end{align*}
\]

\[
\begin{align*}
    &\text{Conv} \\
    &\text{Conv} \\
    &\text{Conv} \\
\end{align*}
\]
Spatial Softmax

Apply Expectation in both axis

\[
f_y = \sum_{ij} s_{ij} = \]

\[
0 \times (0 + 0.66 + 0.03) + \\
1 \times (0 + 0.001 + 0.24) + \\
2 \times (0.09 + 0 + 0.006) = 1.159
\]
Experiments

- The goal
  - Does it work?
  - Can it generalize?
  - Comparison
    - Contextual Policy
    - DA-LSTM Policy
    - DAML, Linear Loss
    - DAML, Temporal Loss
Tasks

- Placing (Meta learning, 1293 of human and robots)
  - Test on 15 different objects
  - One-Shot Learning
    - Placing a held object
    - Avoid two distractors
Tasks

- Pushing (Meta learning, 640 of human and robots)
  - Test on 12 different objects
  - One-Shot Learning
    - Pushing an object
    - Amid one distractor
Tasks

- Pick and Place (Meta learning, 1008 of human and robots)
  - Test on 15 different
  - One-Shot Learning
    - Pick an Place object
    - Avoid two distractors
Some subset of objects

Placing Target

Holding Objects

Pushing Objects

Training  Testing

Training  Testing

Training  Testing

Training  Testing
Experiments

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

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

*denotes equal contribution
Results

Table 1: One-shot success rate of PR2 robot

Demonstration from the same perspective as robot.

<table>
<thead>
<tr>
<th>Method</th>
<th>Placing</th>
<th>Pushing</th>
<th>Pick and Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA-LSTM</td>
<td>33.3%</td>
<td>33.3%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Contextual</td>
<td>36.1%</td>
<td>16.7%</td>
<td>16.7%</td>
</tr>
<tr>
<td>DAML, linear loss</td>
<td>76.7%</td>
<td>27.8%</td>
<td>11.1%</td>
</tr>
<tr>
<td>DAML, temporal loss (ours)</td>
<td>93.8%</td>
<td>88.9%</td>
<td>80.0%</td>
</tr>
</tbody>
</table>
Large Domain Shift

Pushing

- Different camera
- Different room
- Different perspective
- 10 table textures
Large Domain Shift

Pushing
- Meta training
- Meta testing
Large Domain Shift

Pushing

- Meta training
- Meta testing
Experiments

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

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

*denotes equal contribution
## Results

<table>
<thead>
<tr>
<th></th>
<th>Seen background</th>
<th>Novel background 1</th>
<th>Novel background 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAML, temporal loss (ours)</td>
<td>81.8%</td>
<td>66.7%</td>
<td>72.7%</td>
</tr>
</tbody>
</table>

### Failure analysis of DAML

<table>
<thead>
<tr>
<th></th>
<th>Seen bg</th>
<th>Novel bg 1</th>
<th>Novel bg 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># successes</td>
<td>27</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td># failures from task identification</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td># failures from control</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>
Experiments

Sawyer robot

- 18 object held-out
- 3 trials
Experiments

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

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

*denotes equal contribution
Results

Sawyer robot  77.8% placing success rate.

18 held-out with 3 trials
Conclusions

● Learning from human demonstration
  ○ Meta training
  ○ Temporal adaptation loss

● Limitations
  ○ Behavior at meta-test are structurally similar to meta-train
  ○ Lots of data required for meta-training

● Beyond human imitation
  ○ Imitating animal or simulated robots
  ○ Inferring information from out of domain
End of Presentation

THANK YOU