MODERN RL LANDSCAPE: PART I

Scott Niekum
Assistant Professor, Department of Computer Science
The University of Texas at Austin
Distributional RL (Bellemare et al. 2017)

\[ Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E} Q(X', A'). \]

VS.

\[ Z(x, a) \overset{D}{=} R(x, a) + \gamma Z(X', A'). \]

Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy \( \pi \), (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).
Distributional RL (Bellemare et al. 2017)

Figure 5. Intrinsic stochasticity in PONG.
Distributional RL (Bellemare et al. 2017)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>&gt; H.B.</th>
<th>&gt; DQN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN</td>
<td>228%</td>
<td>79%</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>DDQN</td>
<td>307%</td>
<td>118%</td>
<td>33</td>
<td>43</td>
</tr>
<tr>
<td>Duel.</td>
<td>373%</td>
<td>151%</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>Prior.</td>
<td>434%</td>
<td>124%</td>
<td>39</td>
<td>48</td>
</tr>
<tr>
<td>Pr. Duel.</td>
<td>592%</td>
<td>172%</td>
<td>39</td>
<td>44</td>
</tr>
<tr>
<td>C51</td>
<td>701%</td>
<td>178%</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>UNREAL†</td>
<td>880%</td>
<td>250%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Figure 6.* Mean and median scores across 57 Atari games, measured as percentages of human baseline (H.B., Nair et al., 2015).

*Figure 7.* Percentage improvement, per-game, of C51 over Double DQN, computed using van Hasselt et al.’s method.
What is distributional RL doing? (Lyle et al. 2019)

- Reduces chattering?
- Stabilizes updates, handles nonstationarity?
- Good auxiliary task?
What is distributional RL doing? (Lyle et al. 2019)

• Identical expectations computed in most tabular and linear approx cases
• And when predictions are different, actually hurts performance often!
• But usually helps with nonlinear function approximation (e.g. DNN)
• Good auxiliary task for representation learning /regularization?
What is meta-learning?

• If you’ve learned 100 tasks already, can you figure out how to learn more efficiently?
  • Now having multiple tasks is a huge advantage!
• Meta-learning = learning to learn
• In practice, very closely related to multi-task learning
• Many formulations
  • Learning an optimizer
  • Learning an RNN that ingests experience
  • Learning a representation
Why is meta-learning a good idea?

• Deep reinforcement learning, especially model-free, requires a huge number of samples

• If we can *meta-learn* a faster reinforcement learner, we can learn new tasks efficiently!

• What can a *meta-learned* learner do differently?
  • Explore more intelligently
  • Avoid trying actions that are know to be useless
  • Acquire the right features more quickly
Meta-learning with supervised learning

meta-training

meta-testing

training data  test set

image credit: Ravi & Larochelle ‘17

Slide credit: Sergey Levine
Meta-learning with supervised learning

supervised learning: $f(x) \rightarrow y$

input (e.g., image)  output (e.g., label)

supervised meta-learning: $f(D_{\text{train}}, x) \rightarrow y$

training set

- How to read in training set?
  - Many options, RNNs can work
  - More on this later

(few shot) training set

(test input) $x_{\text{test}}$

(test label) $y_{\text{test}}$
RNN-based meta-learning
The meta-learning problem in RL

supervised meta-learning: \( f(D_{\text{train}}, x) \rightarrow y \)

reinforcement meta-learning (for example...): \( f(D_{\text{train}}, s) \rightarrow a \)

\( D_{\text{train}} = \{s_1, a_1, r_1, \ldots, a_N, s_N, r_N\} \)
Back to representations...

is pretraining a type of meta-learning?
better features = faster learning of new task!
Preparing a model for faster learning

\[ \theta \leftarrow \theta + \alpha \nabla_\theta R(\theta) \]

\[ \theta \leftarrow \theta + \alpha \sum_i \nabla_\theta R_i[\theta + \alpha \nabla_\theta R_i(\theta)] \]

Finn et al., “Model-Agnostic Meta-Learning”
Meta-learning summary & open problems

• Meta-learning = learning to learn
• Supervised meta-learning = supervised learning with datapoints that are entire datasets
• RL meta-learning with RNN policies
  • Ingest past experience with RNN
  • Simply run forward pass at test time to “learn”
  • Just contextual policies (no actual learning)
• Model-agnostic meta-learning
  • Use gradient descent (e.g., policy gradient) learning rule
  • Conceptually not that different
  • ...but can accelerate standard RL algorithms (e.g., learn in one iteration of PG)
Meta-learning summary & open problems

- The promise of meta-learning: use past experience to simply acquire a much more efficient deep RL algorithm
- The reality of meta-learning: mostly works well on smaller problems
- ...but getting better all the time
- Main limitations
  - RNN policies are extremely hard to train, and likely not scalable
  - Model-agnostic meta-learning presents a tough optimization problem
  - Designing the right task distribution is hard
  - Generally very sensitive to task distribution (meta-overfitting)
Why not just initialize parameters to those that give the best average performance across tasks?

Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation $\theta$ that can quickly adapt to new tasks.
Isn’t MAML just parameter initialization?

No! Surprisingly, MAML is universal: it can learn any update rule, in principle.
Leveraging auxiliary data sources and multiple data modalities for increased efficiency

Auxiliary video alignment

Natural language narration

Gaze and facial expressions

W. Goo and S. Niekum.
One Shot Learning of Multi-Step Tasks from Observation via Activity Localization in Auxiliary Video

P. Goyal, S. Niekum, and R. Mooney.
Using Natural Language for Reward Shaping in RL
International Joint Conference on AI, August 2019.

A. Saran, E.S. Short, A.L. Thomaz, and S. Niekum.
Understanding Teacher Gaze Patterns for Robot Learning.
Conference on Robot Learning (CoRL), October 2019.

"Jump over the skull while going to the left"
Colored Target Reaching Task

Subtask A: Reaching to an orange target

Subtask B: Reaching to a green target
One-Shot Learning from Observation for Multi-Step Tasks via Activity Localization in Auxiliary Video

Meta-learn a low-shot activity classifier

...then perform IRL
Experiment - Meta-Training

Meta-Training dataset; videos with preselected 36 target colors, 100 videos per each task

$\tau_1$; target orange and green

$\tau_2$; target blue and yellow

$\tau_n$; target purple and red
Learning from Observation (LfO) - Approach

- Learning a notion of *progress*
  - Shuffle-and-Learn loss

Are the frames in order?

\[
g(\text{\begin{array}{c} \circ \end{array}}, \text{\begin{array}{c} \circ \end{array}}) = 1; \text{ in order}
\]

\[
g(\text{\begin{array}{c} \circ \end{array}}, \text{\begin{array}{c} \circ \end{array}}) = 0; \text{ out of order}
\]

For all possible pairs,

\[
\text{Loss} = L_{ce}(\text{sigmoid}(g(o_t, o_{t'}))), 1(t < t')
\]
Learning from Observation (LfO) - Result
Result - the whole pipeline

Graph showing the success rate over 100 trials with different settings:
- Green line with dots: with GT rewards
- Blue line with squares: with Single Demo.
- Orange line with triangles: without Seg.
- Red line with triangles: Ours
- Purple line with circles: with GT Seg.

Y-axis: Success Rate (over 100 Trials)
X-axis: # RL Updates
Results - Breakfast dataset
Gaze – a signal of Human Intent
Gaze Patterns in Human Demonstrations for Robots

Keyframe-based Kinesthetic Teaching (KT)  Observational/Video Demonstrations
Gaze Fixations during Ambiguous Placement Demonstrations

Instruction: Place Green Ladle to the right of Yellow Bowl

More fixations on the yellow bowl
Gaze Fixations during Ambiguous Placement Demonstrations

Instruction: Place Green Ladle to the left of Red Plate

****FIXATION****

Red Plate

More fixations on the red plate
Kinesthetic Demos:
Novice Users focus more on the Robot’s Gripper
Gaze augmented Bayesian IRL for Placement Task

\[ P(R|D, G) \propto P(D|R)P(R|G) \]

Penalize reward functions for which pairwise gaze fixation times are not ranked according to corresponding object weights

Reward functions modeled as weighted RBF kernels near objects
Bayesian IRL using Gaze from Ambiguous Demonstrations

“Place green ladle to the right of the yellow bowl”
“Place green ladle to the right of the yellow bowl”

BIRL without Gaze Information

Proposed ladle location from learnt policy
BIRL with Gaze Information

“Place green ladle to the right of the yellow bowl”
Coverage-based Gaze Loss (CGL)

- Only required during training as part of an auxiliary loss function
- Can be applied to any existing Imitation Learning network with convolutional layers
- Improved performance without varying model complexity

Intuition: Add a penalty for regions where gaze fixations are non-zero, but are not attended to by convolutional layers
CGL: Coverage-based Gaze Loss

(a) Input image    (b) Human    (c) T-REX    (d) T-REX+CGL

A. Saran, R. Zhang, E.S. Short, and S. Niekum. 
Efficiently Guiding Imitation Learning Algorithms with Human Gaze. 
# BCO and T-REX + Gaze

Table 1: BCO performance with and without the usage of human demonstrators’ gaze

<table>
<thead>
<tr>
<th>Game</th>
<th>Human</th>
<th>BCO</th>
<th>BCO+GMD</th>
<th>BCO+CGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakout</td>
<td>344 - 554</td>
<td>0.2</td>
<td>0.0</td>
<td><strong>0.6</strong></td>
</tr>
<tr>
<td>Hero</td>
<td>34305 - 50485</td>
<td>0.0</td>
<td>0.0</td>
<td><strong>1469.0</strong></td>
</tr>
<tr>
<td>MsPacman</td>
<td>17441 - 92610</td>
<td>90.0</td>
<td>70.0</td>
<td><strong>210.0</strong></td>
</tr>
<tr>
<td>Asterix</td>
<td>88000-537500</td>
<td><strong>650.0</strong></td>
<td>363.3</td>
<td>336.7</td>
</tr>
<tr>
<td>Phoenix</td>
<td>22410-27570</td>
<td>24.0</td>
<td>389.3</td>
<td><strong>656.3</strong></td>
</tr>
<tr>
<td>Space Invaders</td>
<td>845-2035</td>
<td>0.0</td>
<td>88.3</td>
<td><strong>311.2</strong></td>
</tr>
<tr>
<td>Enduro</td>
<td>278-742</td>
<td>0.0</td>
<td>0.0</td>
<td><strong>3.2</strong></td>
</tr>
</tbody>
</table>

Table 2: T-REX performance with and without the usage of expert human demonstrators’ gaze

<table>
<thead>
<tr>
<th>Game</th>
<th>Human</th>
<th>T-REX</th>
<th>T-REX+CGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asterix</td>
<td>88000-537500</td>
<td>23926.7</td>
<td><strong>99468.3</strong></td>
</tr>
<tr>
<td>Centipede</td>
<td>39737-251961</td>
<td><strong>12862.8</strong></td>
<td>8514.3</td>
</tr>
<tr>
<td>Phoenix</td>
<td>22410-27570</td>
<td>542.00</td>
<td><strong>669.7</strong></td>
</tr>
<tr>
<td>MsPacman</td>
<td>27731-36061</td>
<td>596.3</td>
<td><strong>625.7</strong></td>
</tr>
</tbody>
</table>
Multimodal data sources: Facial Reactions

Implicit human feedback:

- Occurs naturally
- Is not necessarily intended to influence behavior
- Can be used with no additional burden on user
EMPATHIC: Learning from implicit feedback — training

Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox. 
The EMPATHIC Framework for Task Learning from Implicit Human Feedback. 
Conference on Robot Learning (CoRL), November 2020.
EMPATHIC: Learning from implicit feedback — deployment

Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox.
The EMPATHIC Framework for Task Learning from Implicit Human Feedback.
Conference on Robot Learning (CoRL), November 2020.