Supervised Learning of Behaviors: Deep Learning, Dynamical Systems, and Behavior Cloning

CS 294-112: Deep Reinforcement Learning
Week 2, Lecture 1
Sergey Levine
Today’s Lecture

1. Definition of sequential decision problems
2. Imitation learning: supervised learning for decision making
   a. Does direct imitation work?
   b. How can we make it work more often?
3. Case studies of recent work in (deep) imitation learning
4. What is missing from imitation learning?

• Goals:
  • Understand definitions & notation
  • Understand basic imitation learning algorithms
  • Understand their strengths & weaknesses
Terminology & notation

- $o_t$ - observation
- $x_t$ - state
- $u_t$ - action

$\pi_\theta(u|o_t)$ - policy
Terminology & notation

- $o_t$ – observation
- $u_t$ – action
- $x_t$ – state
- $\pi_\theta(u_t | o_t)$ – policy
- Markov property independent of $x_{t-1}$
Terminology & notation

- $o_t$ – observation
- $x_t$ – state
- $u_t$ – action
- $\pi_\theta(u_t|o_t)$ – policy

A bit of history...
- $x_t$ – state
- $u_t$ – action
- $s_t$ – state
- $a_t$ – action

Lev Pontryagin
Richard Bellman
Imitation Learning

\[ o_t \quad \pi_{\theta}(u_t|o_t) \quad u_t \]

Images: Bojarski et al. ’16, NVIDIA
Does it work?  No!
Does it work? Yes!

Video: Bojarski et al. ‘16, NVIDIA
Why did that work?

Bojarski et al. ‘16, NVIDIA
Can we make it work more often?

stability
Learning from a stabilizing controller

$p(x)$, Gaussian distribution obtained using variant of iterative LQR

(more on this later)
Can we make it work more often?

\[
\pi_\theta(\mathbf{u}_t | \mathbf{o}_t)
\]

can we make \(p_{\text{data}}(\mathbf{o}_t) = p_{\pi_\theta}(\mathbf{o}_t)\)?
Can we make it work more often?

can we make $p_{\text{data}}(o_t) = p_{\pi_\theta}(o_t)$?

idea: instead of being clever about $p_{\pi_\theta}(o_t)$, be clever about $p_{\text{data}}(o_t)$!

**DAgger: Dataset Aggregation**

goal: collect training data from $p_{\pi_\theta}(o_t)$ instead of $p_{\text{data}}(o_t)$

how? just run $\pi_\theta(u_t | o_t)$

but need labels $u_t$!

1. train $\pi_\theta(u_t | o_t)$ from human data $\mathcal{D} = \{o_1, u_1, \ldots, o_N, u_N\}$
2. run $\pi_\theta(u_t | o_t)$ to get dataset $\mathcal{D}_\pi = \{o_1, \ldots, o_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $u_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

Ross et al. ‘11
DAgger Example

Ross et al. ‘11
What’s the problem?

1. train $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$ from human data $\mathcal{D} = \{\mathbf{o}_1, \mathbf{u}_1, \ldots, \mathbf{o}_N, \mathbf{u}_N\}$
2. run $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$ to get dataset $\mathcal{D}_\pi = \{\mathbf{o}_1, \ldots, \mathbf{o}_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $\mathbf{u}_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

Ross et al. ‘11
Imitation learning: recap

- Often (but not always) insufficient by itself
  - Distribution mismatch problem
- Sometimes works well
  - Hacks (e.g. left/right images)
  - Samples from a stable trajectory distribution
  - Add more **on-policy** data, e.g. using DAgger
Case study 1: trail following as classification

A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots

Alessandro Giusti, Jérôme Guzzi, Dan C. Cireşan, Fang-Lin He, Juan P. Rodriguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, Davide Scaramuzza, Luca M. Gambardella
Case study 2: DAgger & domain adaptation

Learning Transferable Policies for Monocular Reactive MAV Control

Shreyansh Daftry, J. Andrew Bagnell, and Martial Hebert

Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
{daftry, dbagnell, hebert}@ri.cmu.edu
1. train $\pi_\theta(u_t|o_t)$ from human data $\mathcal{D} = \{o_1, u_1, \ldots, o_N, u_N\}$
2. run $\pi_\theta(u_t|o_t)$ to get dataset $\mathcal{D}_\pi = \{o_1, \ldots, o_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $u_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
Fig. 2. Experiments and Results for (Row-1) Transfer across physical systems from ARDrone to ArduCopter, (Row-2) Transfer across weather conditions from summer to winter and (Row-3) Transfer across environments from Univ. of Zurich to CMU.
Case study 3: Imitation with LSTMs

Learning real manipulation tasks from virtual demonstrations using LSTM

Rouhollah Rahmatizadeh¹, Pooya Abolghasemi¹, Aman Behal² and Ladislau Bölöni¹
Learning Manipulation Trajectories Using Recurrent Neural Networks
Other topics in imitation learning

• Structured prediction
  
  x: where are you
  y: I’m at work

  ▪ See Mohammad Norouzi’s lecture in April!

• Interaction & active learning

• Inverse reinforcement learning
  
  ▪ Instead of copying the demonstration, figure out the goal
  ▪ Will be covered later in this course
Imitation learning: what’s the problem?

• Humans need to provide data, which is typically finite
  • Deep learning works best when data is plentiful

• Humans are not good at providing some kinds of actions

• Humans can learn autonomously; can our machines do the same?
  • Unlimited data from own experience
  • Continuous self-improvement
Next time: learning without humans
Terminology & notation

- $x_t$ – state
- $o_t$ – observation
- $u_t$ – action

Cost function: $c(x_t, u_t)$
Reward function: $r(x_t, u_t)$

$$\min_{u_1, \ldots, u_T} \sum_{t=1}^{T} \log p(o_t | x_t, u_t) + \lambda \cdot f(u_t)$$
Cost/reward functions in theory and practice

\[ r(x, u) = \begin{cases} 
1 & \text{if object at target} \\
0 & \text{otherwise} 
\end{cases} \]

\[ r(x, u) = -w_1\|p_{\text{gripper}}(x) - p_{\text{object}}(x)\|^2 + \\
- w_2\|p_{\text{object}}(x) - p_{\text{target}}(x)\|^2 + \\
- w_3\|u\|^2 \]

\[ r(x, u) = w_1 v(x) + \\
w_2 \delta(\|\theta_{\text{torso}}(x)\| < \epsilon) + \\
w_3 \delta(h_{\text{torso}}(x) \geq h) \]
A cost function for imitation?

\[ c(x, u) = -\log p(u = \pi^*(x) | x) \]

1. train \( \pi_\theta(u_t|o_t) \) from human data \( \mathcal{D} = \{o_1, u_1, \ldots, o_N, u_N\} \)
2. run \( \pi_\theta(u_t|o_t) \) to get dataset \( \mathcal{D}_\pi = \{o_1, \ldots, o_M\} \)
3. Ask human to label \( \mathcal{D}_\pi \) with actions \( u_t \)
4. Aggregate: \( \mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi \)

Ross et al. '11
The trouble with cost & reward functions

reinforcement learning agent

what is the reward?

Sim-to-Real Robot Learning from Pixels with Progressive Nets

Andrei A. Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, Raia Hadsell

Google DeepMind
London, UK

{andreirusu, matejvecerik, tcr, heess, razp, raia}@google.com

Rewards are given automatically by tracking the colored target

More on this later...
A note about terminology...
the “R” word

a bit of history...
reinforcement learning (the problem statement)

\[ \min \sum_{t=1}^{T} E[c(x_t, u_t)] \quad x_{t+1} \sim p(x_{t+1} | x_t, u_t) \]

reinforcement learning (the method)
without using the model

\[ x_{t+1} \sim p(x_{t+1} | x_t, u_t) \]

Lev Pontryagin  Richard Bellman  Andrew Barto  Richard Sutton