Deep Reinforcement Learning

Xiyuan Yang
2025.11.22

Lecture Notes for Deep Reinforcement Learning, CS285

目录

§1. Introduction

§1.1. Works to Cover

1. From supervised learning to decision making
2. Model-free algorithms: Q-learning, policy gradients, actor-critic
3. Model-based algorithms: planning, sequence models, etc.
4. Exploration
5. Offline reinforcement learning
6. Inverse reinforcement learning
7. Advanced topics, research talks, and invited lectures

§1.2. Introduction to RL

§1.2.1. Supervised Learning

Given $D = \{(x_i, y_i)\}$, we want the supervised learning systems to learn how to predict y from x: $f(x) \approx y$.

It usually assumes:

- i.i.d data.(独立同分布)
- known ground truth outputs in training

For example, Deep Learning for Image Recognitions/Classifications. (**Need High-Labeled Data**)

§1.2.2. Reinforcement Learning

- Data is not i.i.d: previous outputs influence
- Ground truth answer is not known, only know

if we succeeded or failed, more generally, we know the reward

> **Recordings.**
> 强化学习对数据的利用更加的松弛，不需要高质量人工标注的数据，这提升了强化学习的上限，但是这也导致模型对数据的利用率较低。

In the mathematical view:

- goal for supervised learning: $f_{\theta(x_i)} = y_i$
  - training data $\{(x_i, y_i)\}$ are fixed and manually labeled.
- goal for reinforcement learning: learning $\pi_\theta : s_t \to a_t$ to maximize $\sum_t r_t$
  - the data $(s_1, a_1, r_1, ..., s_T, a_T, r_T)$: own actions, dynamic!

§1.2.2.1. Applications

- games, robotics
- RL with Large Language Models
- RL with image generations
- RL for chip design

§1.2.3. Deep Reinforcement Learning

Supervised Learning has the upper-bound has labeled data, but RL does not.

"Move 37" in Lee Sedol AlphaGo match: reinforcement learning "discovers" a move that surprises everyone.

- Data Driven AI (learns about the real world from data, but doesn't try to do better than the data)
- Reinforcement Learning (optimizes a goal with emergent behavior, but need to figure out how to use **at scale**).

Combination: **Deep Reinforcement Learning**!

> **Recordings** The Bitter Lesson.
> We have to learn the bitter lesson that building in how we think we think does not work in the long run.

- Data without optimization doesn't allow us to solve new problems in new ways.
- Optimization without data is hard to apply to the real world **outside of simulators**.

The core components (two general building blocks for AI-systems):

- Learning: use data to extract patterns (world laws), understanding the world
- Search: Use computations to extract inferences. Making inferences and leverages that understanding for emergence.

> **Recordings.**
> We have a brain for one reason and one reason only – that's to produce adaptable and complex movements. Movement is the only way we have affecting the world around us··· I believe that to understand movement is to understand the whole brain.

§1.2.4. Sequential Decision Making

Far more than a convex optimization problem!

- Learning reward functions from example (inverse reinforcement learning)
- Transferring knowledge between domains (transfer learning, meta-learning)
- Learning to predict and using prediction to act

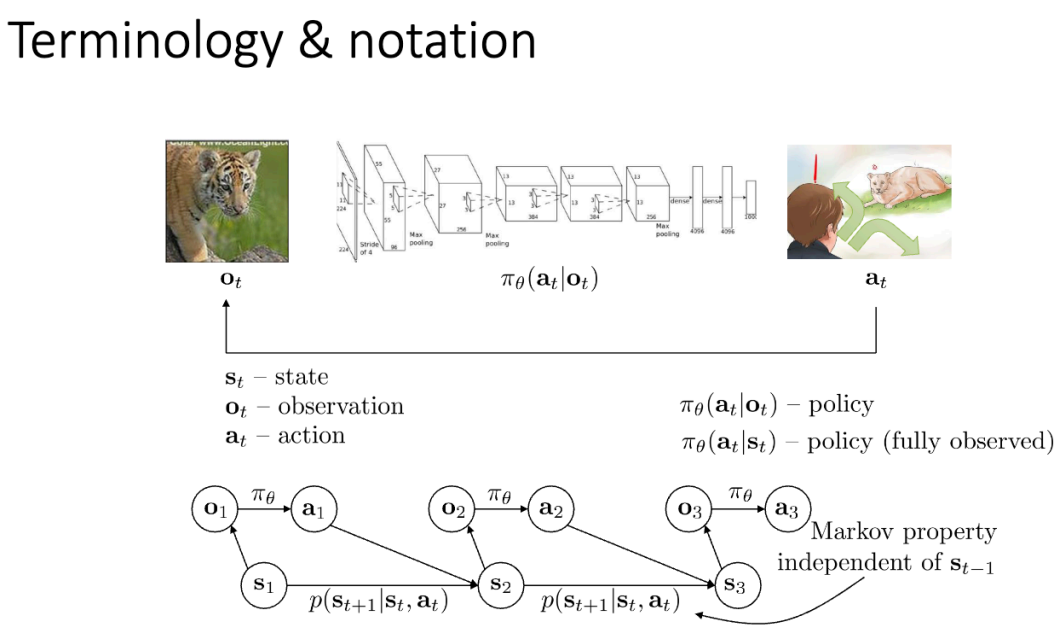
real-time reward is hard to design.

- Learning from demonstrations
  - Directly copying observed behavior
  - Inferring rewards from observed behavior (inverse reinforcement learning)
- Learning from observing the world
  - Learning to predict
  - Unsupervised learning
- Learning from other tasks
  - Transfer learning
  - Meta-learning: learning to learn

Will RL be the way to AGI? (using a **general learning algorithms** for interacting observations and actions with the environment)

§1.3. Supervised Learning of Behaviors

- policy based on observations: $\pi_\theta(a_t|o_t)$
- policy based on full observations: $\pi_\theta(a_t|s_t)$
- policy are distributions (probability)



§2. Conclusion