A Beginner's Guide to Deep Reinforcement Learning

When it is not in our power to determine what is true, we ought to act in accordance with what is most probable. - Descartes

Contents

Introduction

Deep reinforcement learning combines artificial neural networks with a framework of reinforcement learning that helps software agents learn how to reach their goals. That is, it unites function approximation and target optimization, mapping states and actions to the rewards they lead to.

You may not understand all of those terms, but they will be explained below, in greater depth and plainer language, drawing on your personal experiences as an individual moving through the world.

While neural networks are responsible for recent AI breakthroughs in problems like computer vision, machine translation and time series prediction – they can also combine with reinforcement learning algorithms to create something astounding like Deepmind's AlphaGo (https://deepmind.com/blog/alphago-zero-learning-scratch/), an algorithm that beat the world champions of the Go board game. That's why you should care about deep RL.

Reinforcement learning refers to goal-oriented algorithms, which learn how to achieve a complex objective (goal) or how to maximize along a particular dimension over many steps; for example, they can maximize the points won in a game over many moves. Reinforcement learning algorithms can start from a blank slate, and under the right conditions, achieve superhuman performance. Like a pet incentivized by scolding and treats, these algorithms are penalized when they make the wrong decisions and rewarded when they make the right ones – this is reinforcement.

Reinforcement algorithms that incorporate deep neural networks can beat human experts playing numerous Atari video games (https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf), Starcraft II (https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning) and Dota-2 (https://openai.com/projects/five/). While that may

sound trivial to non-gamers, it's a vast improvement over reinforcement learning's previous accomplishments, and the state of the art is progressing rapidly.

Reinforcement learning solves the hard problem of correlating immediate actions with the delayed outcomes they produce. Like humans, reinforcement learning algorithms sometimes have to wait to see the fruit of their decisions. They operate in a delayed-return environment, where it can be difficult to understand which action leads to which outcome over many time steps.

Reinforcement learning algorithms are slowly performing better and better in more ambiguous, real-life environments while choosing from an arbitrary number of possible actions, rather than from the limited options of a repeatable video game. That is, they are beginning to achieve goals in the real world. If you have measurable KPIs to reach, deep RL may be able to help. DeepMind claimed in May 2021 (https://www.sciencedirect.com/science/article/pii/S0004370221000862) that reinforcement learning was probably sufficient to achieve artificial general intelligence (AGI) (/strong-ai-general-ai).

Companies are beginning to apply deep reinforcement learning to problems in industry.

Pieter Abbeel's Covariant (https://covariant.ai/) uses deep RL in industrial robotics. Pathmind applies deep reinforcement learning to simulations (https://pathmind.com/) of industrial operations and supply chains to optimize factories, warehouses and logistics. Google is applying deep RL to problems such as robot locomotion and chip design (https://ai.googleblog.com/2020/08/a-simulation-suite-for-tackling-applied.htmlhttps://ai.googleblog.com/2020/08/a-simulation-suite-for-tackling-applied.html), while Microsoft relies on deep RL to power its autonomous control systems technology (https://www.microsoft.com/en-us/ai/autonomous-systems).

Reinforcement Learning Definitions

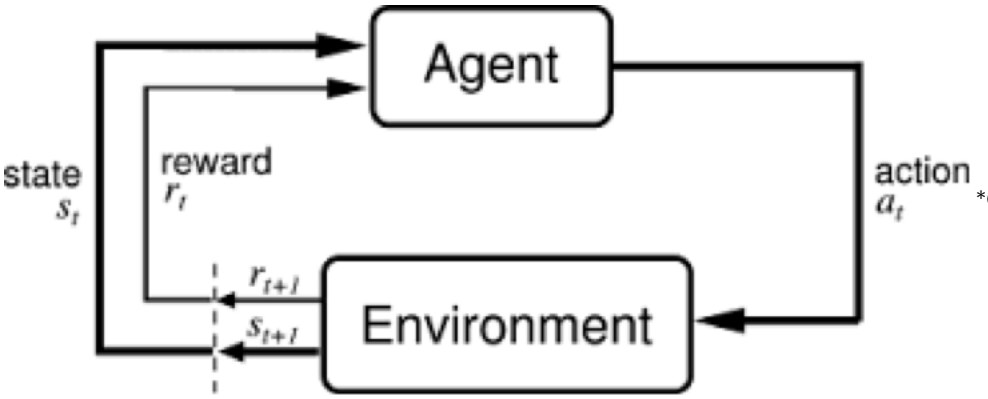
Let's learn the basic terms of RL.

Reinforcement learning can be understood through the concepts of agents, environments, states, actions and rewards, all of which we'll explain below. Capital letters tend to denote sets of things, and lower-case letters denote a specific instance of that thing; e.g. `A` is all possible actions, while `a` is a specific action contained in the set.

- Agent: An **agent** takes actions; for example, a drone making a delivery, or Super Mario navigating a video game. The algorithm is the agent. It may be helpful to consider that in life, the agent is you.[1]
- Action (A): `A` is the set of all possible moves the agent can make. An **action** is almost self-explanatory, but it should be noted that agents usually choose from a list of discrete, possible actions. In video games, the list might include running right or left, jumping high or low, crouching or standing still. In the stock markets, the list might include buying, selling or holding any one of an array of securities and their derivatives. When handling aerial drones, alternatives would include many different velocities and accelerations in 3D space.
- Discount factor: The **discount factor** is multiplied by future rewards as discovered by the agent in order to dampen thse rewards' effect on the agent's choice of action. Why? It is designed to make future rewards worth less than immediate rewards; i.e. it enforces a kind of short-term hedonism in the agent. Often expressed with the lower-case Greek letter gamma: *γ*. If *γ* is .8, and there's a reward of 10 points after 3 time steps, the present value of that reward is `0.8³ x 10`. A discount factor of 1 would make future rewards worth just as much as immediate rewards. We're fighting against delayed gratification (https://en.wikipedia.org/wiki/Stanford_marshmallow_experiment) here.
- Environment: The world through which the agent moves, and which responds to the agent. The environment takes the agent's current state and action as input, and returns as output the agent's reward and its next state. If you are the agent, the environment could be the laws of physics and the rules of society that process your actions and determine the consequences of them.
- State (S): A **state** is a concrete and immediate situation in which the agent finds itself; i.e. a specific place and moment, an instantaneous configuration that puts the agent in relation to other significant things such as tools, obstacles, enemies or prizes. It can the current situation returned by the environment, or any future situation. Were you ever in the wrong place at the wrong time? That's a state.
- Reward (R): A **reward** is the feedback by which we measure the success or failure of an agent's actions in a given state. For example, in a video game, when Mario touches a coin, he wins points. From any given state, an agent sends output in the form of actions to the environment, and the environment returns the agent's new state (which resulted from acting on the previous state) as well as rewards, if there are any. Rewards can be immediate or delayed. They effectively evaluate the agent's action.
- Policy (π): The **policy** is the strategy that the agent employs to determine the next action based on the current state. It maps states to actions, the actions that promise the highest reward.
- Value (V): The expected long-term return with discount, as opposed to the short-term reward `R`. `Vπ(s)` is defined as the expected long-term return of the current state under policy `π`. We discount rewards, or lower their estimated value, the further into the future they occur. See *discount factor*. And remember Keynes: "In the long run, we are all dead." That's why you discount future rewards. It is useful to distinguish

- Q-value or action-value (Q): **Q-value** is similar to Value, except that it takes an extra parameter, the current action `a`. `Qπ(s, a)` refers to the long-term return of an action taking action a under policy `π` from the current state `s`. Q maps state-action pairs to rewards. Note the difference between Q and policy.
- Trajectory: A sequence of states and actions that influence those states. From the Latin "to throw across." The life of an agent is but a ball tossed high and arching through space-time unmoored, much like humans in the modern world.
- Key distinctions: Reward is an immediate signal that is received in a given state, while value is the sum of all rewards you might anticipate from that state. Value is a long-term expectation, while reward is an immediate pleasure. Value is eating spinach salad for dinner in anticipation of a long and healthy life; reward is eating cocaine for dinner and to hell with it. They differ in their time horizons. So you can have states where value and reward diverge: you might receive a low, immediate reward (spinach) even as you move to position with great potential for long-term value; or you might receive a high immediate reward (cocaine) that leads to diminishing prospects over time. This is why the value function, rather than immediate rewards, is what reinforcement learning seeks to predict and control.

So environments are functions that transform an action taken in the current state into the next state and a reward; agents are functions that transform the new state and reward into the next action. We can know and set the agent's function, but in most situations where it is useful and interesting to apply reinforcement learning, we do not know the function of the environment. It is a black box where we only see the inputs and outputs. It's like most people's relationship with technology: we know what it does, but we don't know how it works. Reinforcement learning represents an agent's attempt to approximate the environment's function, such that we can send actions into the black-box environment that maximize the rewards it spits out.



*Credit: Sutton & Barto

(http://incompleteideas.net/book/bookdraft2017nov5.pdf)

In the feedback loop above, the subscripts denote the time steps `t` and `t+1`, each of which refer to different states: the state at moment `t`, and the state at moment `t+1`. Unlike other forms of machine learning – such as supervised and unsupervised learning – reinforcement learning can only be thought about sequentially in terms of state-action pairs that occur one after the other.

Reinforcement learning judges actions by the results they produce. It is goal oriented, and its aim is to learn sequences of actions that will lead an agent to achieve its goal, or maximize its objective function. Here are some examples:

- In video games, the goal is to finish the game with the most points, so each additional point obtained throughout the game will affect the agent's subsequent behavior; i.e. the agent may learn that it should shoot battleships, touch coins or dodge meteors to maximize its score.
- In the real world, the goal might be for a robot to travel from point A to point B, and every inch the robot is able to move closer to point B could be counted like points.

Here's an example of an objective function for reinforcement learning; i.e. the way it defines its goal.

$$\sum_{t=0}^{t=\infty} \gamma^t r(x(t), a(t))$$

We are summing reward function *r* over *t*, which stands for time steps. So this objective function calculates all the reward we could obtain by running through, say, a game. Here, *x* is the state at a given time step, and *a* is the action taken in that state. *r* is the reward function for *x* and *a*. (We'll ignore *γ* for now.)

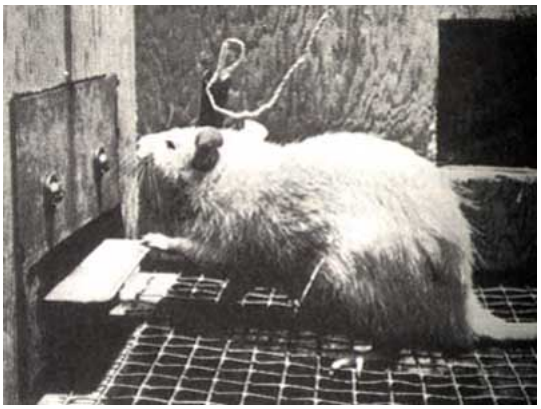
Reinforcement learning differs from both supervised and unsupervised learning by how it interprets inputs. We can illustrate their difference by describing what they learn about a "thing."

- Unsupervised learning: That thing is like this other thing. (The algorithms learn similarities w/o names, and by extension they can spot the inverse and perform anomaly detection by recognizing what is unusual or dissimilar)
- Supervised learning: That thing is a "double bacon cheese burger". (Labels, putting names to faces…) These algorithms learn the correlations between data instances and their labels; that is, they require a labelled dataset. Those labels are used to "supervise" and correct the algorithm as it makes wrong guesses when predicting labels.
- Reinforcement learning: Eat that thing because it tastes good and will keep you alive longer. (Actions based on short- and long-term rewards, such as the amount of calories you ingest, or the length of time you survive.) Reinforcement learning can be thought of as supervised learning in an environment of sparse feedback.

Domain Selection for Reinforcement Learning

One way to imagine an autonomous reinforcement learning agent would be as a blind person attempting to navigate the world with only their ears and a white cane. Agents have small windows that allow them to perceive their environment, and those windows may not even be the most appropriate way for them to perceive what's around them.

(In fact, deciding *which types* of input and feedback your agent should pay attention to is a hard problem to solve. This is known as domain selection. Algorithms that are learning how to play video games can mostly ignore this problem, since the environment is man-made and strictly limited. Thus, video games provide the sterile environment of the lab, where ideas about reinforcement learning can be tested. Domain selection requires human decisions, usually based on knowledge or theories about the problem to be solved; e.g. selecting the domain of input for an algorithm in a self-driving car might include choosing to include radar sensors in addition to cameras and GPS data.)



State-Action Pairs & Complex Probability Distributions of Reward

The goal of reinforcement learning is to pick the best known action for any given state, which means the actions have to be ranked, and assigned values relative to one another. Since those actions are state-dependent, what we are really gauging is the value of state-action pairs; i.e. an action taken from a certain state, something you did somewhere. Here are a few examples to demonstrate that the value and meaning of an action is contingent upon the state in which it is taken:

- If the action is marrying someone, then marrying a 35-year-old when you're 18 probably means something different than marrying a 35-year-old when you're 90, and those two outcomes probably have different motivations and lead to different outcomes.
- If the action is yelling "Fire!", then performing the action a crowded theater should mean something different from performing the action next to a squad of men with rifles. We can't predict an action's outcome without knowing the context.

We map state-action pairs to the values we expect them to produce with the Q function, described above. The Q function takes as its input an agent's state and action, and maps them to probable rewards.

Reinforcement learning is the process of running the agent through sequences of state-action pairs, observing the rewards that result, and adapting the predictions of the Q function to those rewards until it accurately predicts the best path for the agent to take. That prediction is known as a policy.

Reinforcement learning is an attempt to model a complex probability distribution of rewards in relation to a very large number of state-action pairs. This is one reason reinforcement learning is paired with, say, a Markov decision process (./markov-chain-monte-carlo), a method to sample from a complex distribution to infer its properties. It closely resembles the problem that inspired Stan Ulam to invent the Monte Carlo method (http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-88-9068); namely, trying to infer the chances that a given hand of solitaire will turn out successful.

Any statistical approach is essentially a confession of ignorance. The immense complexity of some phenomena (biological, political, sociological, or related to board games) make it impossible to reason from first principles. The only way to study them is through statistics, measuring superficial events and attempting to establish correlations between them, even when we do not understand the mechanism by which they relate. Reinforcement learning, like deep neural networks, is one such strategy, relying on sampling to extract information from data.

After a little time spent employing something like a Markov decision process to approximate the probability distribution of reward over state-action pairs, a reinforcement learning algorithm may tend to repeat actions that lead to reward and cease to test alternatives. There is a tension between the exploitation of known rewards, and continued exploration to discover new actions that also lead to victory. Just as oil companies have the dual function of pumping crude out of known oil fields while drilling for new reserves, so too, reinforcement learning algorithms can be made to both **exploit** and **explore** to varying degrees, in order to ensure that they don't pass over rewarding actions at the expense of known winners.

Reinforcement learning is iterative. In its most interesting applications, it doesn't begin by knowing which rewards state-action pairs will produce. It learns those relations by running through states again and again, like athletes or musicians iterate through states in an attempt to improve their performance.

The Relationship Between Machine Learning with Time

You could say that an algorithm is a method to more quickly aggregate the lessons of time.[2] Reinforcement learning algorithms have a different relationship to time than humans do. An algorithm can run through the same states over and over again while experimenting with different actions, until it can infer which actions are best from which states. Effectively, algorithms enjoy their very own Groundhog Day (https://www.imdb.com/title/tt0107048/), where they start out as dumb jerks and slowly get wise.

Since humans never experience Groundhog Day outside the movie, reinforcement learning algorithms have the potential to learn more, and better, than humans. Indeed, the true advantage of these algorithms over humans stems not so much from their inherent nature, but from their ability to live in parallel on many chips at once, to train night and day without fatigue, and therefore to learn more. An algorithm trained on the game of Go, such as AlphaGo, will have played many more games of Go than any human could hope to complete in 100 lifetimes.[3]

Neural Networks and Deep Reinforcement Learning

Where do neural networks fit in?

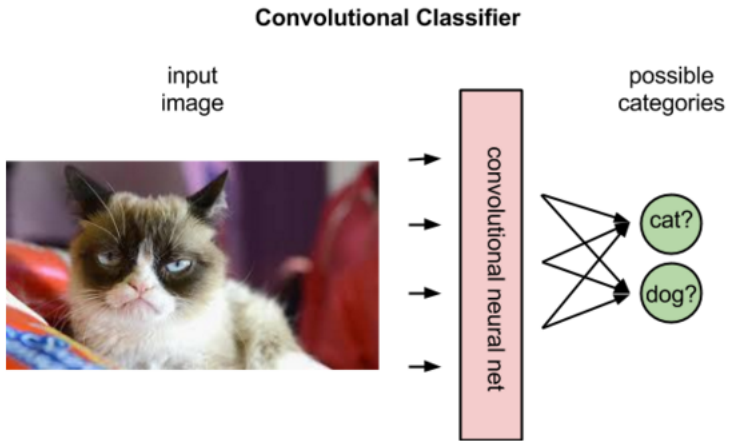
Neural networks (./neural-network) are function approximators, which are particularly useful in reinforcement learning when the state space or action space are too large to be completely known.

A neural network can be used to approximate a value function, or a policy function. That is, neural nets can learn to map states to values, or state-action pairs to Q values. Rather than use a lookup table to store, index and update all possible states and their values, which impossible with very large problems, we can train a neural network on samples from the state or action space to learn to predict how valuable those are relative to our target in reinforcement learning.

Like all neural networks, they use coefficients to approximate the function relating inputs to outputs, and their learning consists to finding the right coefficients, or weights, by iteratively adjusting those weights along gradients that promise less error.

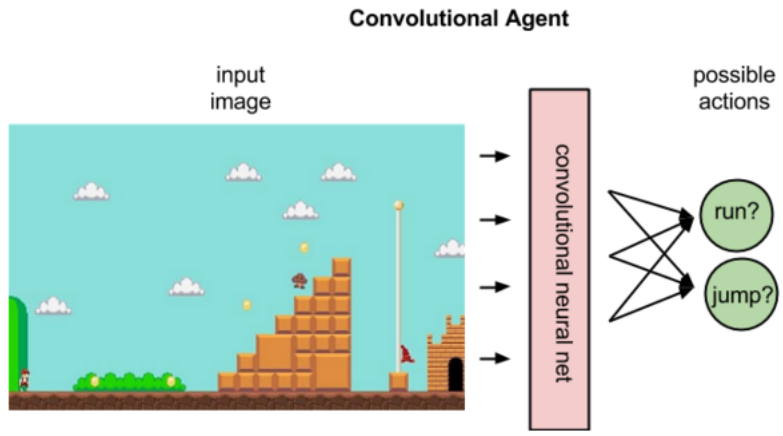
In reinforcement learning, convolutional networks can be used to recognize an agent's state when the input is visual; e.g. the screen that Mario is on, or the terrain before a drone. That is, they perform their typical task of image recognition.

But convolutional networks derive different interpretations from images in reinforcement learning than in supervised learning. In supervised learning, the network applies a label to an image; that is, it matches names to pixels.



In fact, it will rank the labels that best fit the image in terms of their probabilities. Shown an image of a donkey, it might decide the picture is 80% likely to be a donkey, 50% likely to be a horse, and 30% likely to be a dog.

In reinforcement learning, given an image that represents a state, a convolutional net can rank the actions possible to perform in that state; for example, it might predict that running right will return 5 points, jumping 7, and running left none.



The above image illustrates what a policy agent does, mapping a state to the best action.

$$a = \pi(s)$$

A policy maps a state to an action.

If you recall, this is distinct from Q, which maps state action pairs to rewards.

To be more specific, Q maps state-action pairs to the highest combination of immediate reward with all future rewards that might be harvested by later actions in the trajectory. Here is the equation for Q, from Wikipedia:

$$Q(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

Having assigned values to the expected rewards, the Q function simply selects the state-action pair with the highest so-called Q value.

At the beginning of reinforcement learning, the neural network coefficients may be initialized stochastically, or randomly. Using feedback from the environment, the neural net can use the difference between its expected reward and the ground-truth reward to adjust its weights and improve its interpretation of state-action pairs.

This feedback loop is analogous to the backpropagation of error in supervised learning. However, supervised learning begins with knowledge of the ground-truth labels the neural network is trying to predict. Its goal is to create a model that maps different images to their respective names.

Reinforcement learning relies on the environment to send it a scalar number in response to each new action. The rewards returned by the environment can be varied, delayed or affected by unknown variables, introducing noise to the feedback loop.

This leads us to a more complete expression of the Q function, which takes into account not only the immediate rewards produced by an action, but also the delayed rewards that may be returned several time steps deeper in the sequence.

Like human beings, the Q function is recursive. Just as calling the wetware method `human()` contains within it another method `human()`, of which we are all the fruit, calling the Q function on a given state-action pair requires us to call a nested Q function to predict the value of the next state, which in turn depends on the Q function of the state after that, and so forth.

Real-World Reinforcement Learning Applications

(This section is a WIP.)

Reinforcement learning is about making sequential decisions to attain a goal over many steps. While other types of AI perform what you might call perceptive tasks, like recognizing the content of an image, reinforcement learning performs tactical and strategic tasks. Games are a good proxy for problems that reinforcement learning can solve, but RL is also being applied to real-world processes in the private and public sectors.

- Robotics
- Industrial Operations
- Supply Chain & Logistics
- Traffic Control
- Bidding & Advertising
- Recommender Systems
- Load Balancing
- Augmented NLP

Footnotes

1) *It might be helpful to imagine a reinforcement learning algorithm in action, to paint it visually. Let's say the algorithm is learning to play the video game Super Mario. It's trying to get Mario through the game and acquire the most points. To do that, we can spin up lots of different Marios in parallel and run them through the space of all possible game states. It's as though you have 1,000 Marios all tunnelling through a mountain, and as they dig (e.g. as they decide again and again which action to take to affect the game environment), their experience-tunnels branch like the intricate and fractal twigs of a tree. The Marios' experience-tunnels are corridors of light cutting through the mountain. And as in life itself, one successful action may make it more likely that successful action is possible in a larger decision flow, propelling the winning Marios onward. You might also imagine, if each Mario is an agent, that in front of him is a heat map tracking the rewards he can associate with state-action pairs. (Imagine each state-action pair as have its own screen overlayed with heat from yellow to red. The many screens are*