# Policy Networks vs Value Networks in Reinforcement Learning

SAGAR SHARMA Aug 5, 2018 · 4 min read

In Reinforcement Learning, the agents take random decisions in their environment and learns on selecting the right one out of many to achieve their goal and play at a super-human level. Policy and Value Networks are used together in algorithms like Monte Carlo Tree Search to perform Reinforcement Learning. Both the networks are an integral part of a method called Exploration in MCTS algorithm.

They are also known as policy iteration & value iteration since they are calculated many times making it an iterative process.

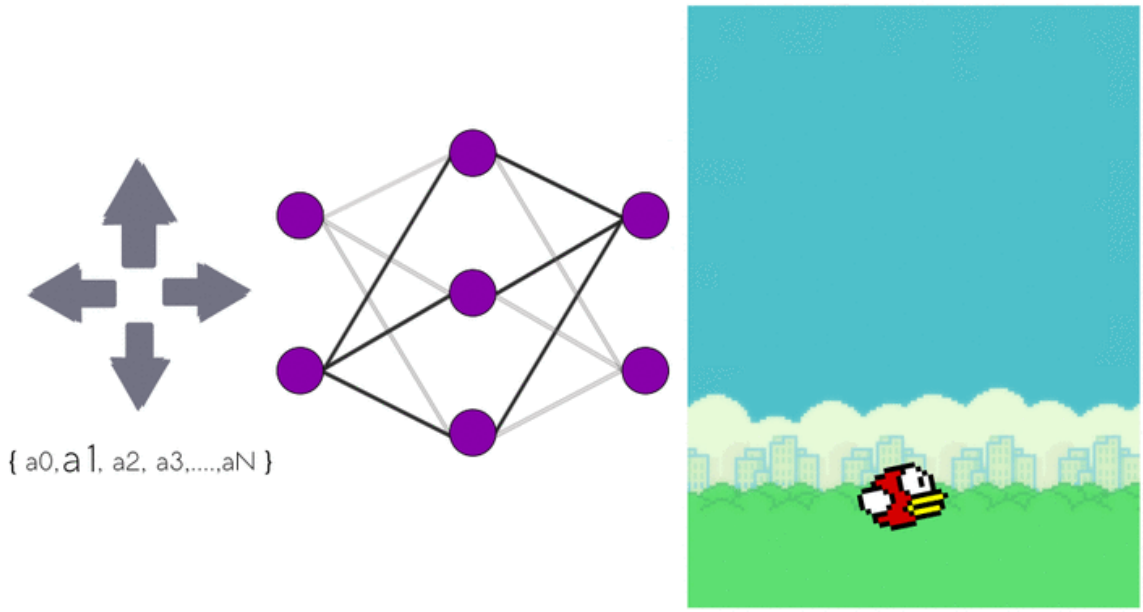
**Let's understand why are they so important in Machine Learning and what's the difference between them?**

## What is a Policy Network?

Consider any game in the world, input 🎮 given by user to the game is known as **actions** $a$ . Every input (action) leads to a different output. These output are known as **states** $s$ of the game.

From this, we can make different state-action pairs $S = \{(s0,a0),s1,a1),...,(sN,aN)\}$ , representing which actions $aN$ leads to which states $sN$. Also, we can say that **S** contains all the policies learned by the policy network.

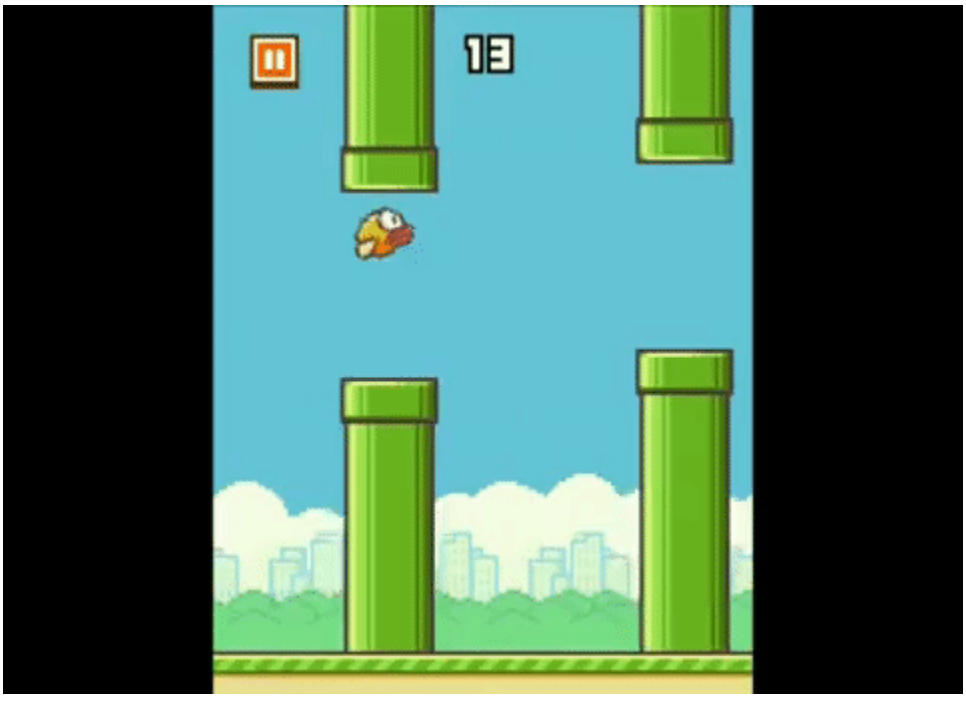
by giving a particular input to the game is known as Policy Network




**Policy Network (**action1, state1**) , (**action2, state2**)**

For Example: Input `a1` gives a state `s1` (moving up) & Input `a2` gives a state `s2` (going down) in the game.

Also, Some actions increase the points of the player lead to reward **r** .



Let's look at some obvious symbols:

$\mathcal{S}$ : set of possible states
$\mathcal{A}$ : set of possible actions
$\mathcal{R}$ : distribution of reward given (state, action) pair
$\mathbb{P}$ : transition probability i.e. distribution over next state given (state, action) pair
$\gamma$ : discount factor

**Usual Notations for RL environments**

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r_t | \pi\right]$$

**Optimal Policy**

### Why are we using Discount Factor $\gamma$

> *It is used as a precautionary measure (usually kept below 1). It prevent the reward* ***r*** *to reach infinite.*

An infinite reward for a policy will overwhelm our agent & biased towards that specific action, killing the desire to explore unknown areas and actions of the game😳.

---

### But how do we know which state to choose for your next move, eventually leading to the final round?

## What is a Value Network?

> *The value network assigns value/score to the state of the game by calculating an expected cumulative score for the current state* `s` *. Every state goes through the value network. The states which gets more reward obviously get more value in the network.*

Keep in mind that the reward is **expected rewards,** because we are choosing the right one from the set of states.



Value Function

Now, the key objective is always to maximise the reward *(aka Markov Decision Process)*. Actions that results in a good state obviously get greater reward than others.

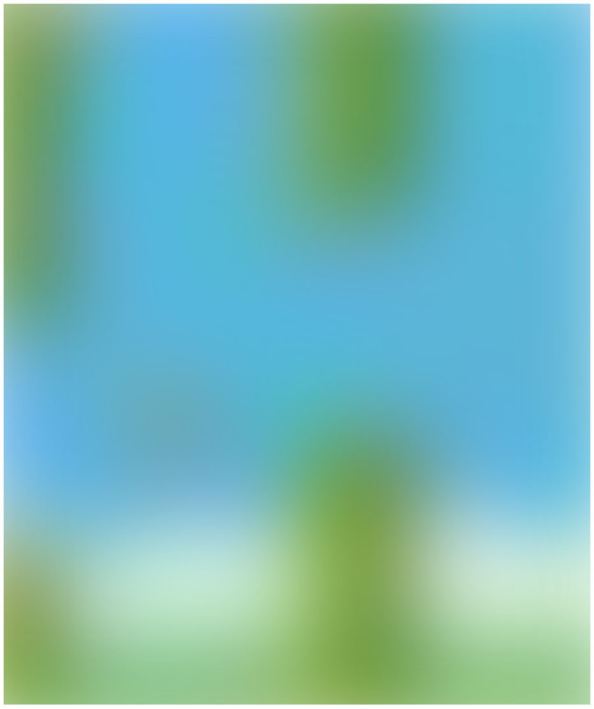
Since any game is won by following a sequence of actions one after the other. The optimal policy $\pi^*$ of the game consists a number of state-action pairs that helps in winning the game.

The state-action pair that achieve most reward is considered as optimal policy.

The equation for optimal policy is formally written using *arg max* as:

**Optimal Policy** $\pi^*$

The optimal policy learned by the policy network knows which actions should be performed at the current state to get maximum reward.



If you have any doubt, query or demand, comment down below or tweet me.

**Clap it… Share it! Follow me** on **Medium** to get similar fun content.

To get instant notification follow me on **Twitter**.


Happy to be helpful. kudos.

## Previous Stories You will Love:

**Monte Carlo Tree Search**
MCTS For Every Data Science Enthusiast
towardsdatascience.com

**TensorFlow Image Recognition Python API Tutorial**
On CPU with Inception-v3(In seconds)
towardsdatascience.com

**Activation Functions: Neural Networks**
Sigmoid, tanh, Softmax, ReLU, Leaky ReLU EXPLAINED !!!
towardsdatascience.com

**DeepMind's Playing Capture The Flag with Deep Reinforcement**

towardsdatascience.com