Why Reinforcement Learning and not Supervised Learning?

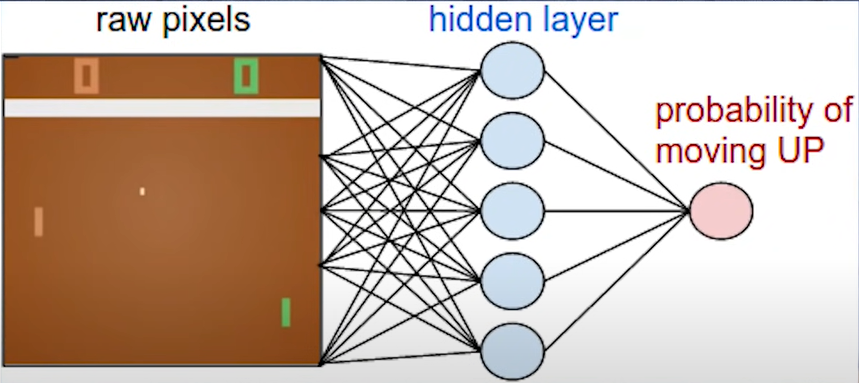
Have to create a very huge dataset to train the model

By Supervised Learning our model will be only able to imitate the user experience but never will be better

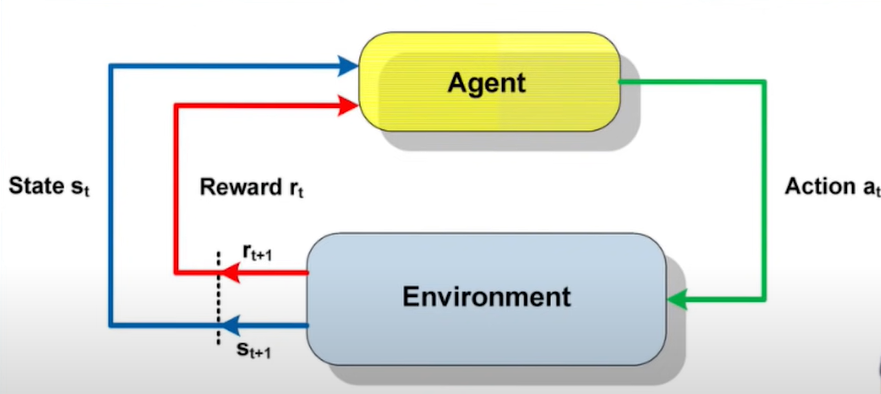
Consider a game of ping pong:

Here if we use supervised learning, we will store frame by frame dataset, that user was doing in that frame(is he pressing up or down arrow). Then we will the train the model on this, that for each type of frames what our model should do

But creating this kind of dataset is a very cumbersome process, also the model will only be able to replicate that humans method of playing the game, it will never be better



Reinforcement Learning:

The network that transforms input frames to output actions is called policy network

To train this policy network we use a method called Policy Gradients.

Over here we do not have a dataset as said above.

So we start out with a random frame from the game, depending on that our policy network will give an output either up or down, then the game engine will give the next frame depending on this output, and so the loop continues

To learn we will be giving our agent a scoreboard, a **+1 score** when our agent scores and a **-1 penalty** when the game engine scores. The entire goal of the agentis to optimize its policy to receive as much reward as possible.

To train, we add a bunch of experience, i.e.we let a whole bunch of games go through our network, select random actions and feed them back into the engine.

For every game we get a positive reward, we use normal gradients to increase probability of those actions in future.

For every penalty we apply the same gradient but with a -1, so all the actions in those games are less likely to happen.

Slowly slowly actions going to penalty will be filtered out and actions giving reward are going to be more likely.

Consider a case where the game was going pretty good and our network was almost going to win, but in the end in the last step it made a mistake, so now the problem here is that the model assumes that all the actions we did in this game are wrong and will give -1 penalty to everything, reducing likelihood of taking those actions in future. This is called the **Credit Assignment Problem.**

Credit Assignment Problem: If we get a reward at the end of our episode, what are the actions that led to that specific reward

Our agent needs to figure what parts of its action sequence are causing the reward that it eventually gets**(Sparse Reward Setting)**. Like here, our agent should learn that the actions just before hitting the ball are really important

**Reward Shaping:** Manually designing a reward function that needs to guide your policy to some desired behavior

**The Alignment Problem:** In a lot of cases when we shape our reward function, the network will find some very surprising way to get the reward but not doing what we wanted it to do, i.e. it will overfit to this reward shaping but not generalizing to the intended behaviour

References:

1. <https://towardsdatascience.com/intro-to-reinforcement-learning-pong-92a94aa0f84d>
2. Introduction to RL: <https://www.youtube.com/watch?v=JgvyzIkgxF0>
3. RL with Sparse Rewards: <https://www.youtube.com/watch?v=0Ey02HT_1Ho>
4. <http://karpathy.github.io/2016/05/31/rl/>