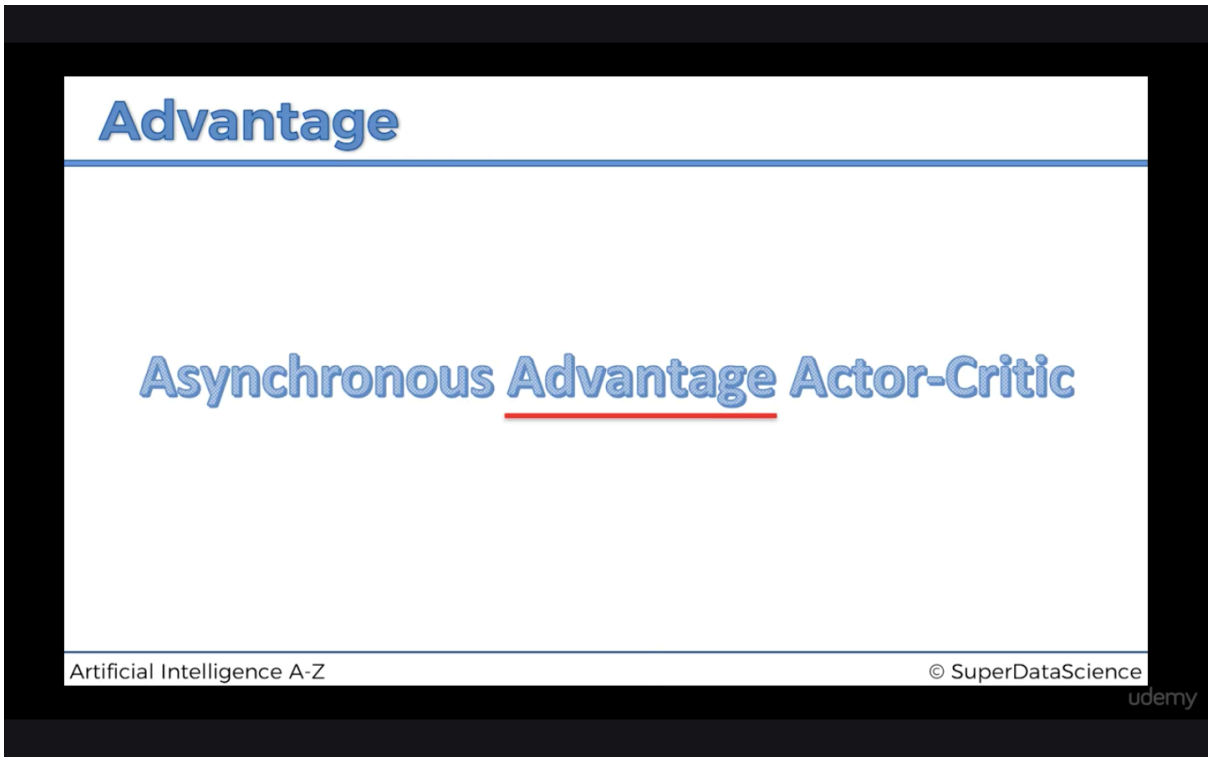
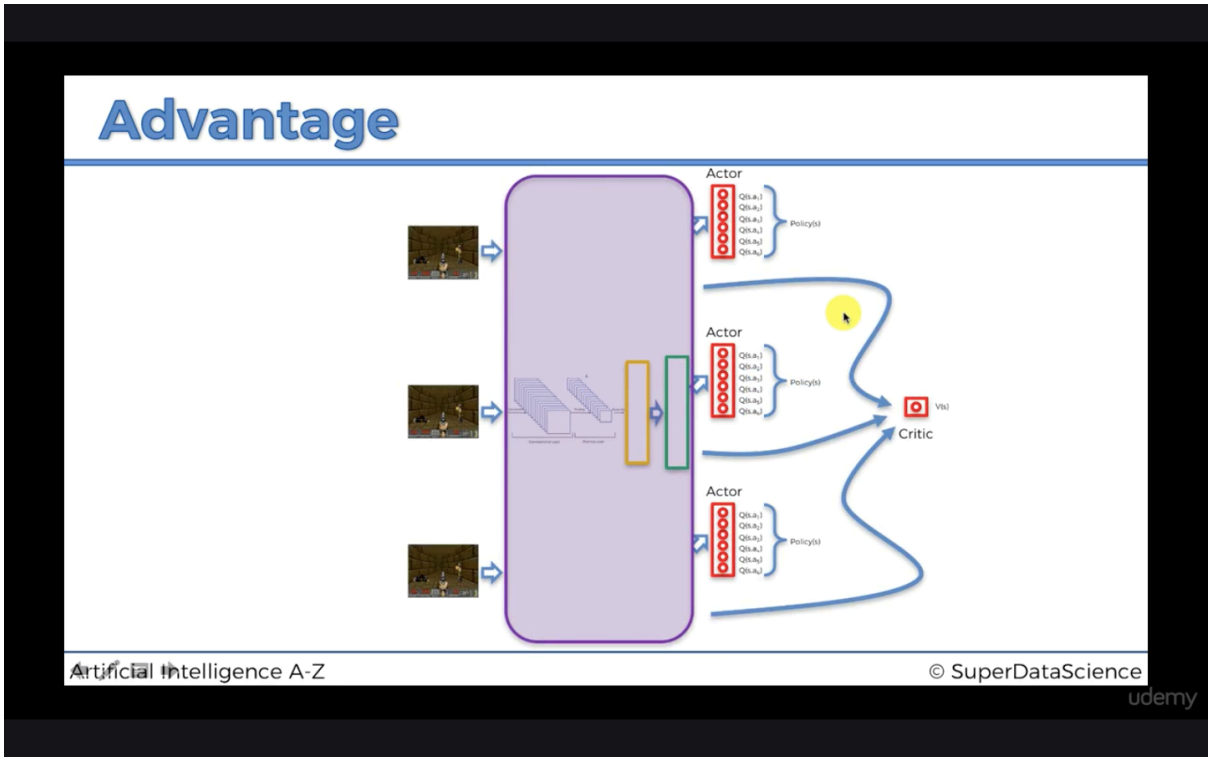
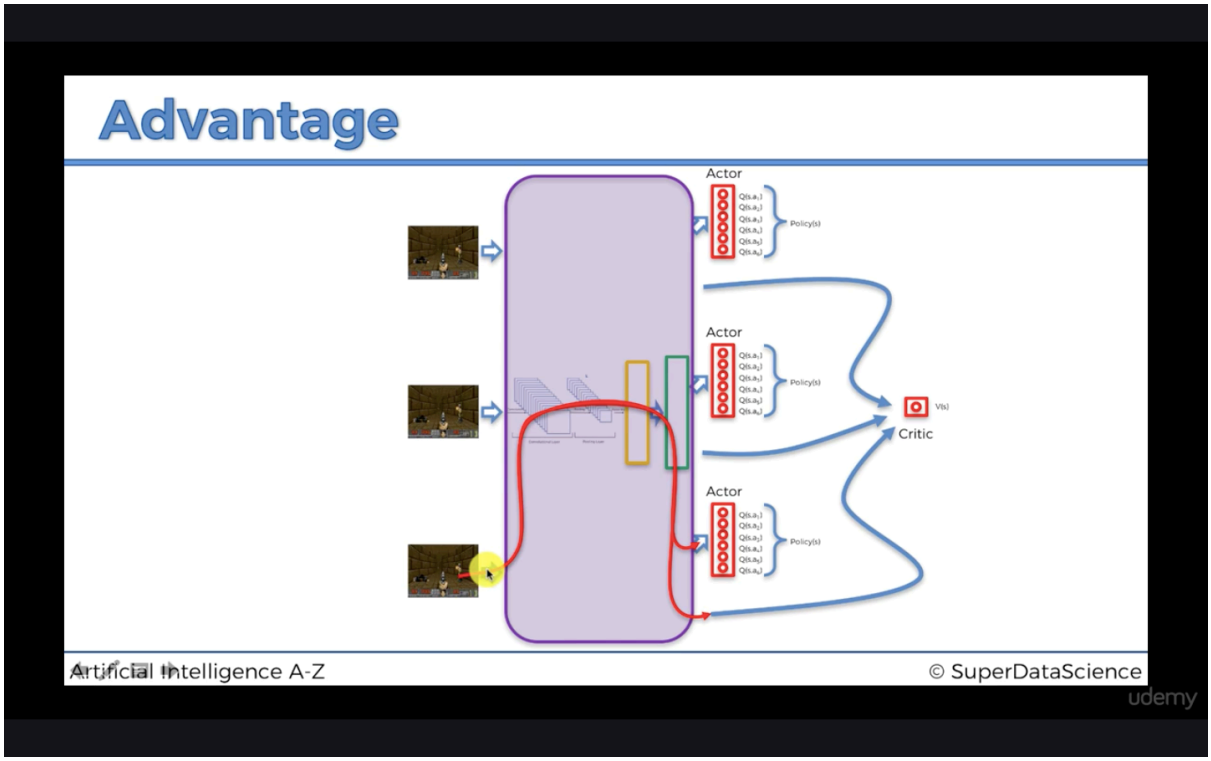
Advantage


Advantage
Artificial Intelligence A-Z
© SuperDataScience
udemy

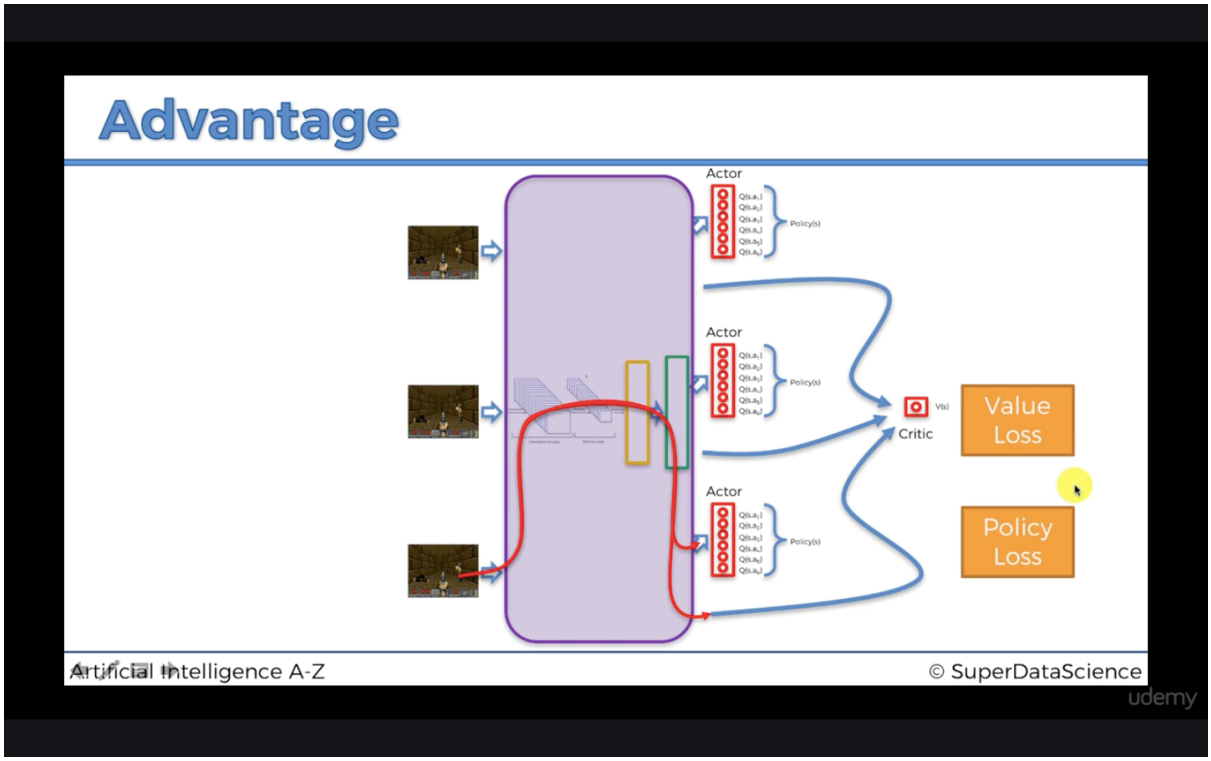

Advantage
Asynchronous Advantage Actor-Critic
Artificial Intelligence A-Z
© SuperDataScience
udemy



Neural network and critic is shared with agents.



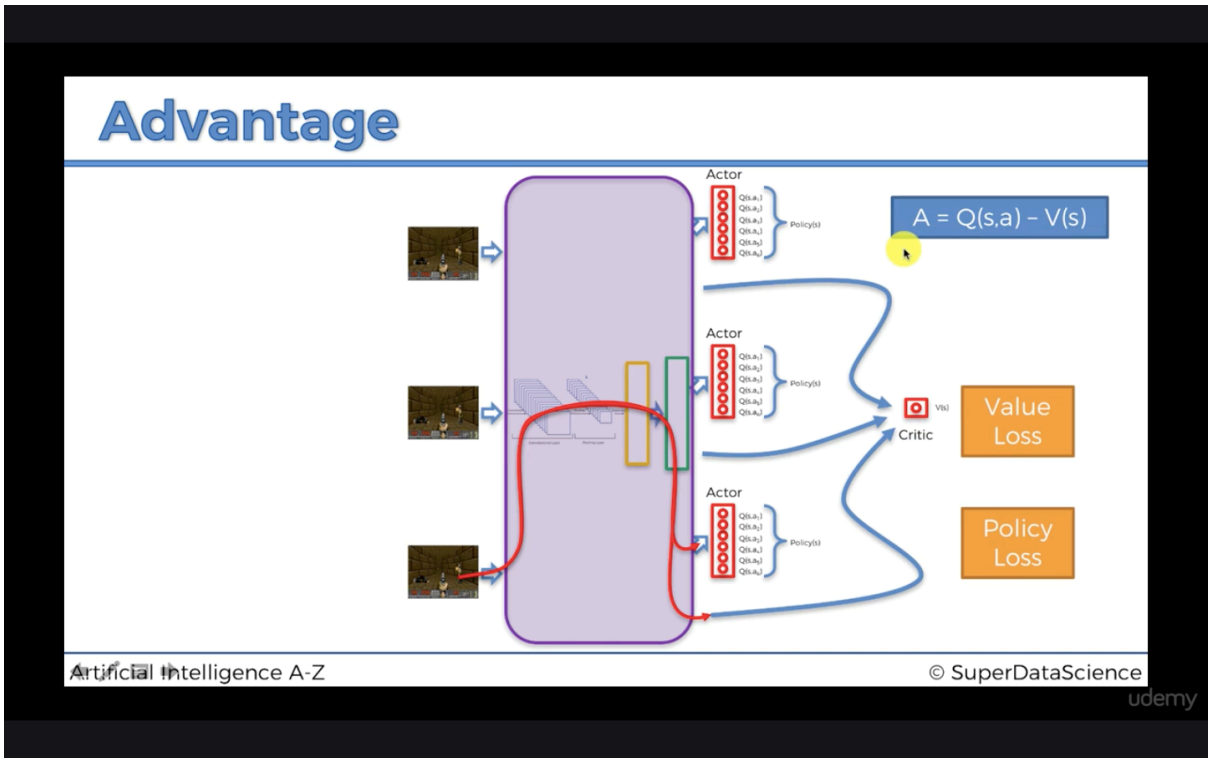
This is when the third agent is at a certain state which want to make a decision to what action to play.

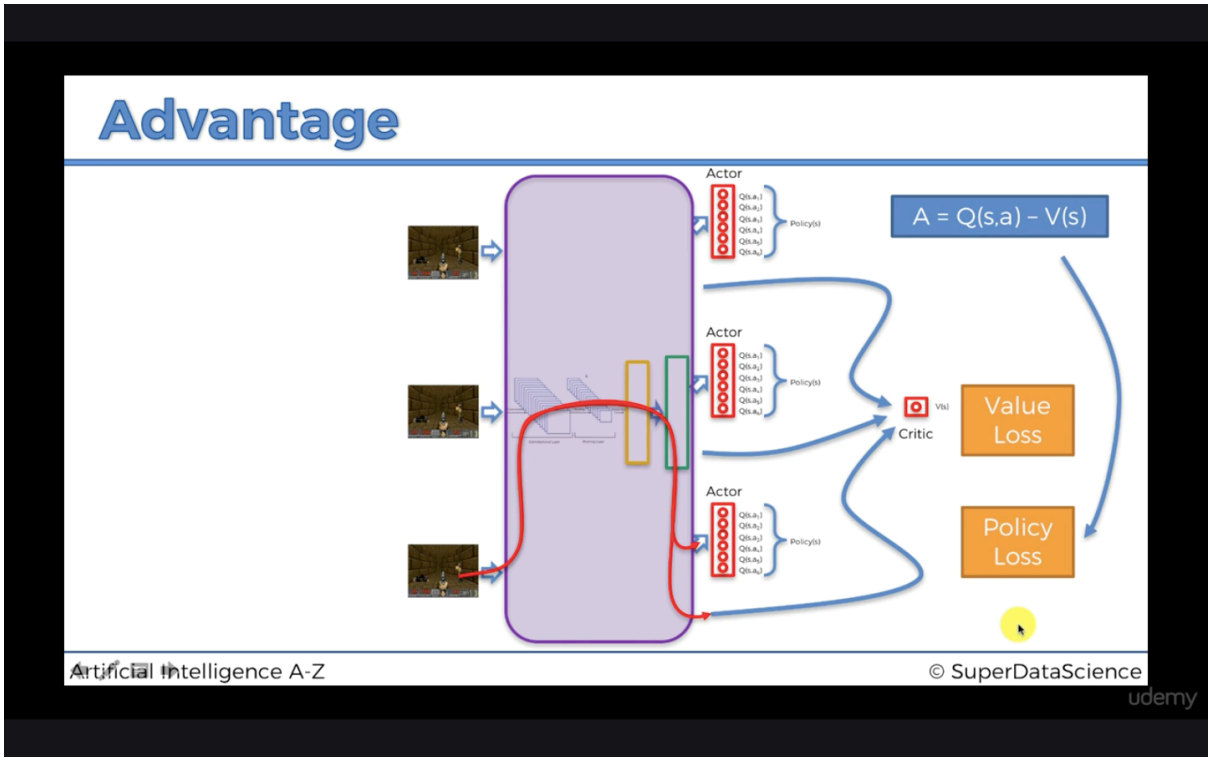


For propagation and calculating the loss, there is two loss which one is for the policy and the other one is for the value we had.

Value loss: the network predicts a certain value V and at the same time we can estimate what should be, based on what we know about the environment so far, we can estimate what should be the value V be in the state and by comparing the two we can calculate the loss and then back propagate.

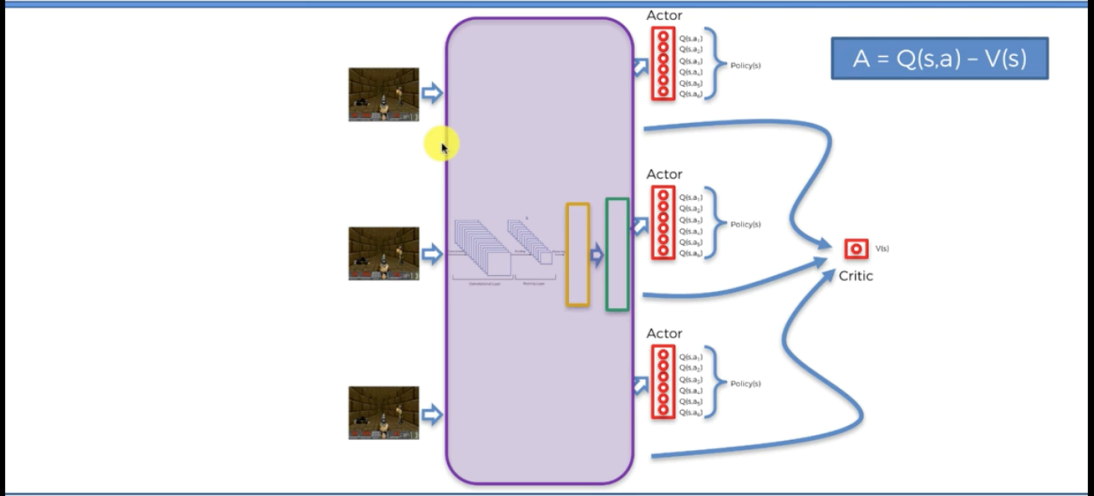
Policy loss: This is when the critic that shared with agent is going to emerge. For calculating the policy loss we need to about value called advantage...



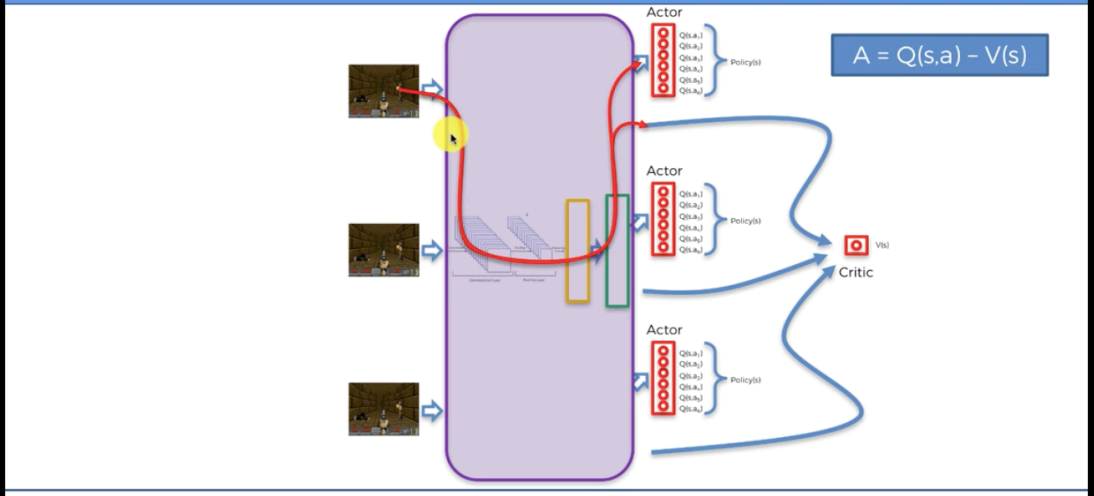


Q-values come from what the neural network predicts for this agent for (s, a) which is the action we choose to play in that state.
V value is dictated by the critic which is the shared part and it can tell us what is the overall known value of that state that these whole agents are performing together.
So, the above advantage formula, critic knows the V(s) value and the question is how much better is the Q-value that you're selecting compared to the know v value. Or in other word what is the advantage that the selected action brings compared to the known value of that state. This helps to calculate the policy loss which it then will back propagate through the network in order to help to maximize the advantage A.
If a bad action (Q-value less than the V) has been chosen then after comparison it will learn that it shouldn't do any of that anymore and the weights are just in a way so that happens is less rare. so that's a less frequent occurrence that we choose for that bad action.
On the other hand, if a good action (Q-value more than the V) has been chosen. After back propagating the network and the weights are going to be updated in such a way to reinforce that it happens again and after A3C algorithm sees the high advantage it wants to more of that again. So the weights are going to be update in such a way that it will be more likely to happen in the future.
Note that we usually choose the Q-values with the highest value. If even after choosing the high value the A was higher, then the weights are going to be update in such a way that that encourages for better actions. On the other hand if we select something that advantage is going to be high then it's going to the policy loss and the network going to be updated.
So the policy loss will encourages more to choose the good choices than making the bad choices.

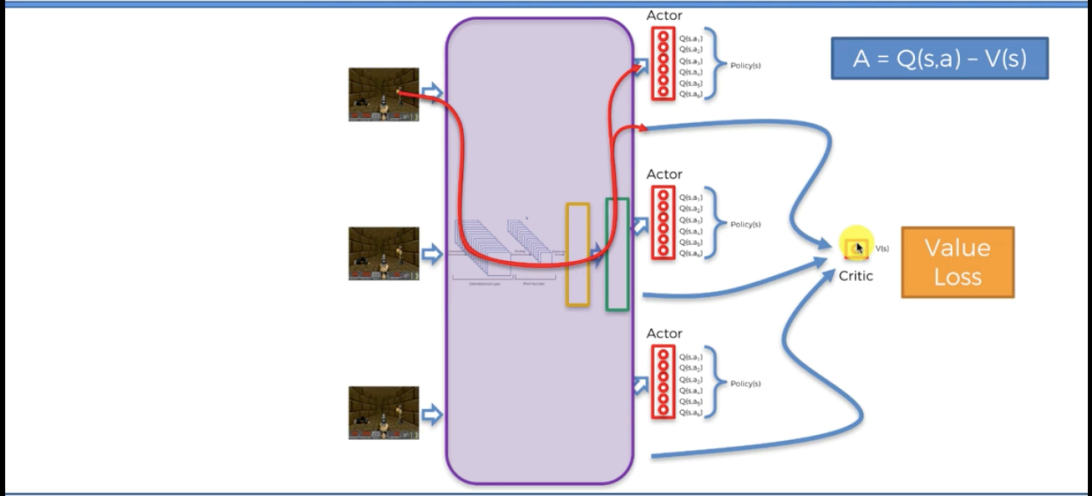
Advantage



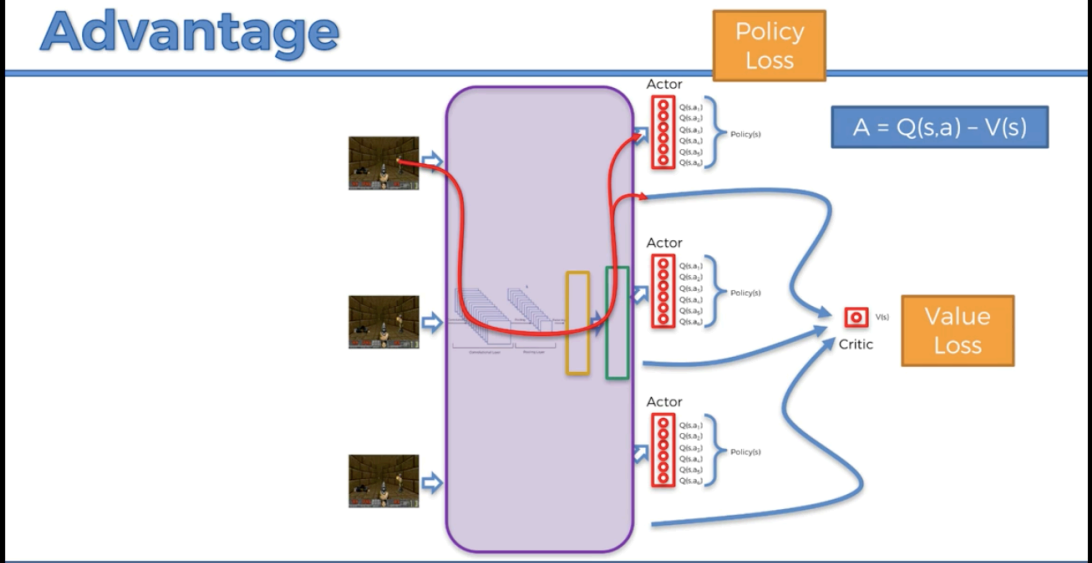
Advantage



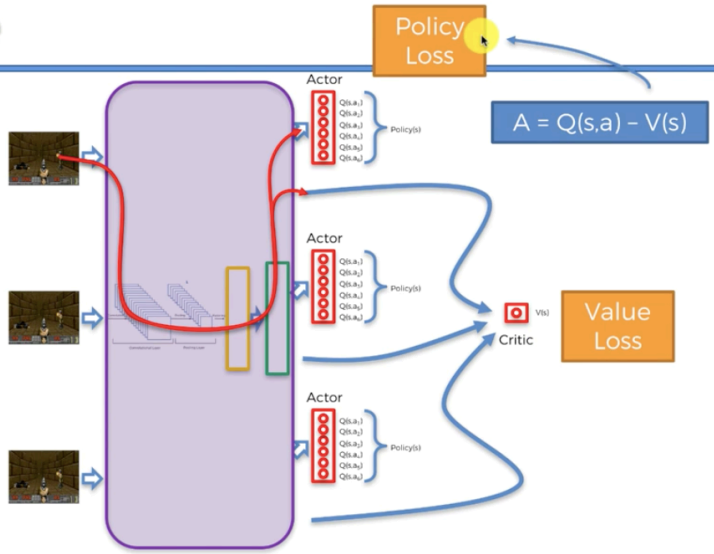
Advantage



Advantage



Advantage



Additional Reading

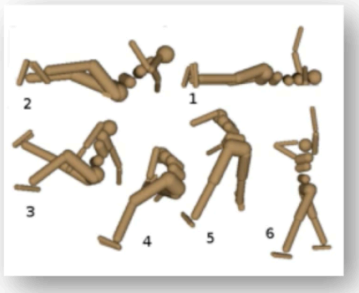
Additional Reading:

High-dimensional Continuous Control Using Generalized Advantage Estimation

John Schulman et al. (2016)

Link:

https://arxiv.org/pdf/1506.02438.pdf

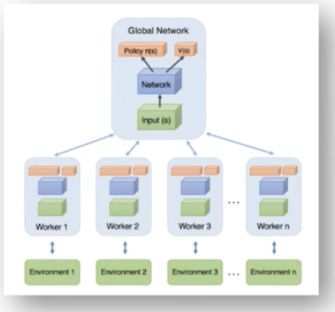


Additional Reading

Additional Reading:

Simple Reinforcement Learning with Tensorflow (Part 8)

By Arthur Juliani (2016)



Link:

https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2