Reinforcement Learning Training 2025

Policy Gradient Algorithm

Concept

- Policy-based methods parameterize the policy (often with a neural network)
- Focus on optimizing the policy itself (not q or v).
- Outputs a probability distribution over actions for each state.

Why?

- Direct policy optimization enables smoother changes
  - Avoiding abrupt behavior shifts seen in value-based methods.
- Can learn stochastic policies natively,
- Handles continuous and large action spaces more naturally

REINFORCE Algorithm

- Learn a policy directly by optimizing the parameters of a policy model using a gradient ascent approach.

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(a_t^i|s_t^i\right)\right)\left(\sum_{t=1}^{T} r\left(s_t^i, a_t^i\right)\right)\right]$$

$$\theta = \theta + \alpha \cdot \nabla_\theta J(\theta)$$

REINFORCE

Input:

A model with parameters θ taking state s as input and producing $\pi_\theta\left(a|s\right)$

Other parameters: step size α

Initialize:

Initialize weights $\boldsymbol{\theta}$

Loop:

Sample $\{\tau^i\}$, a set of N trajectories from current policy $\pi_\theta\left(a_t|s_t\right)$

Update model parameters θ:

$$\nabla_\theta J\left(\theta\right) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(a_t^i|s_t^i\right)\right)\left(\sum_{t=1}^{T} r\left(s_t^i, a_t^i\right)\right)\right]$$

$$\theta = \theta + \alpha\nabla_\theta J\left(\theta\right)$$

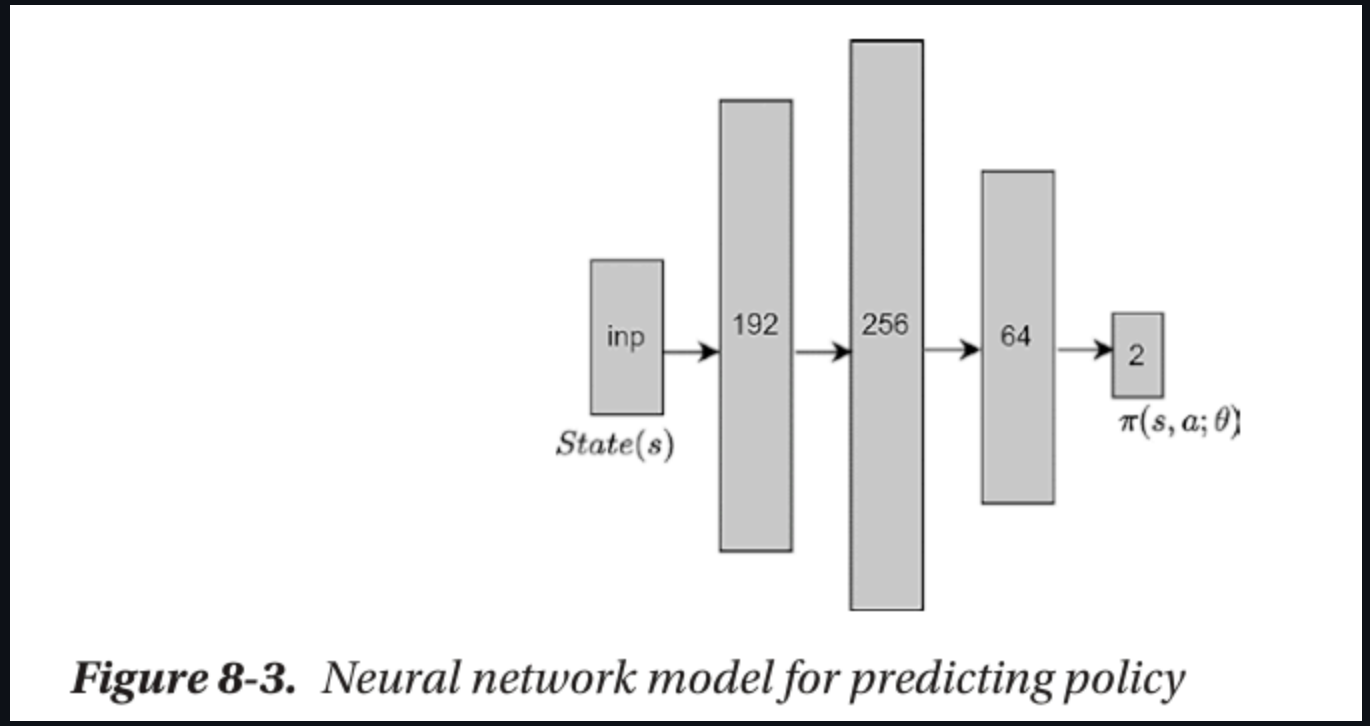

Figure 8-3. *Neural network model for predicting policy*

Actor-Critic

Actor-Critic is a class of reinforcement learning algorithms that combines the strengths of both policy-based and value-based methods by maintaining two separate models: an actor and a critic.

Components

- **Actor**
  - The actor is responsible for selecting actions according to a parameterized policy.
  - It learns to maximize the expected reward by adjusting the policy parameters based on feedback from the critic.
- **Critic**
  - The critic estimates the value function (v or q).
  - Evaluating how good the actions taken by the actor are in terms of expected cumulative rewards.

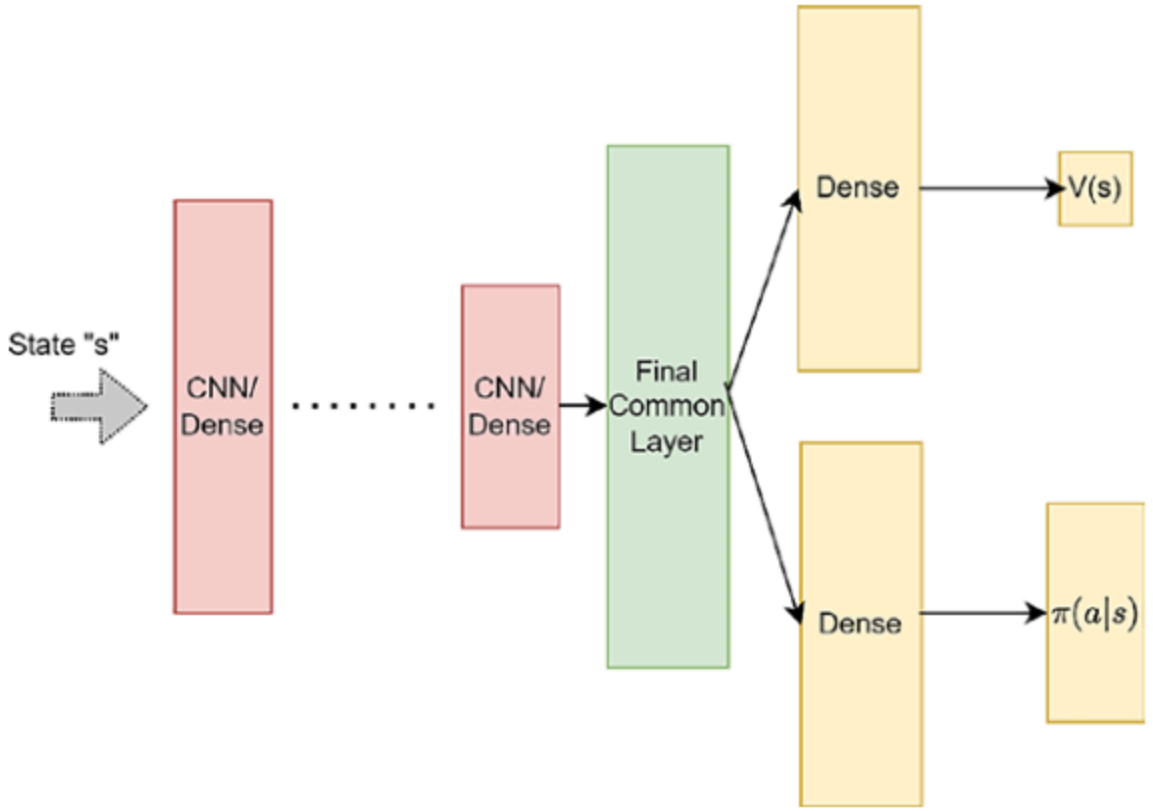

Figure 8-8. *Actor-critic network with common weights in the initial layers*