In this project, we follow the solution detail step by step. We first write mlp function to construct a basic neural network structure. Then we complete the Actor class which takes input as the state observations and returns a probability distribution of the possible action space given this state and current policy. At the class VPGBuffer, we implement to calculate the discounted reward-to-go and discounted advantage function in order to construct a policy gradient estimator. Then in the Agent class, it acts like a combination of actor and critics which returns an action according to a sample from the policy distribution. Eventually, at the training function, both the actor loss and critic loss are defined and optimized based on normal procedure.