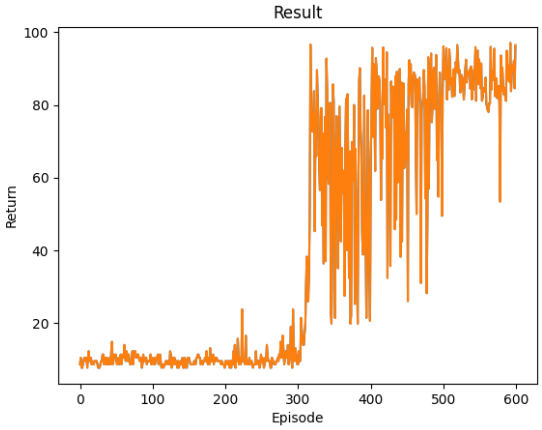

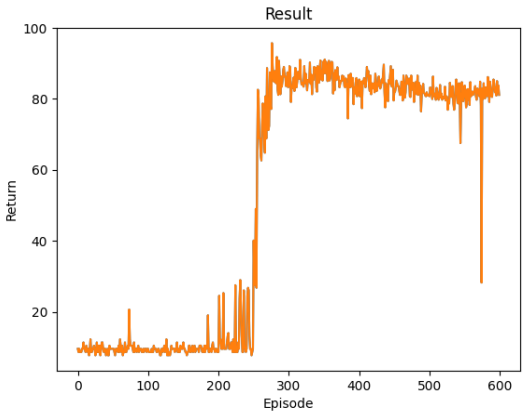

The graph on the left displays a target net strategy while the graph on the right displays a soft update strategy. As gamma is 0.99 and the reward per step never exceeds 1.0, we can expect the return to asymptotically approach near 100, but never increase past this due to gamma. The target net strategy has a clear jump in returns every 100 episodes. As the target network is updated only every 100 episodes, we can see a major change in the returns between episodes 200-300 and episodes 300-400, where the accumulated policy update causes significant differences in the return as the agent is now able to maintain the pole on the cart for a long period of time. The soft update strategy displays much faster learning as learning happens at every timestep. This model is able to achieve returns comparable to the target net strategy at only 300 vs 600 episodes.