Training DQN

- DQN uses experience replay and fixed Q-targets
- Store transition $(s_t, a_t, r_{t+1}, s_{t+1})$ in replay memory D
- Sample random mini-batch of transitions (s, a, r, s') from D
- Compute Q-learning targets w.r.t. old, fixed parameters $\boldsymbol{\theta}^-$
- Optimizes MSE between Q-network and Q-learning targets
- Uses stochastic gradient descent