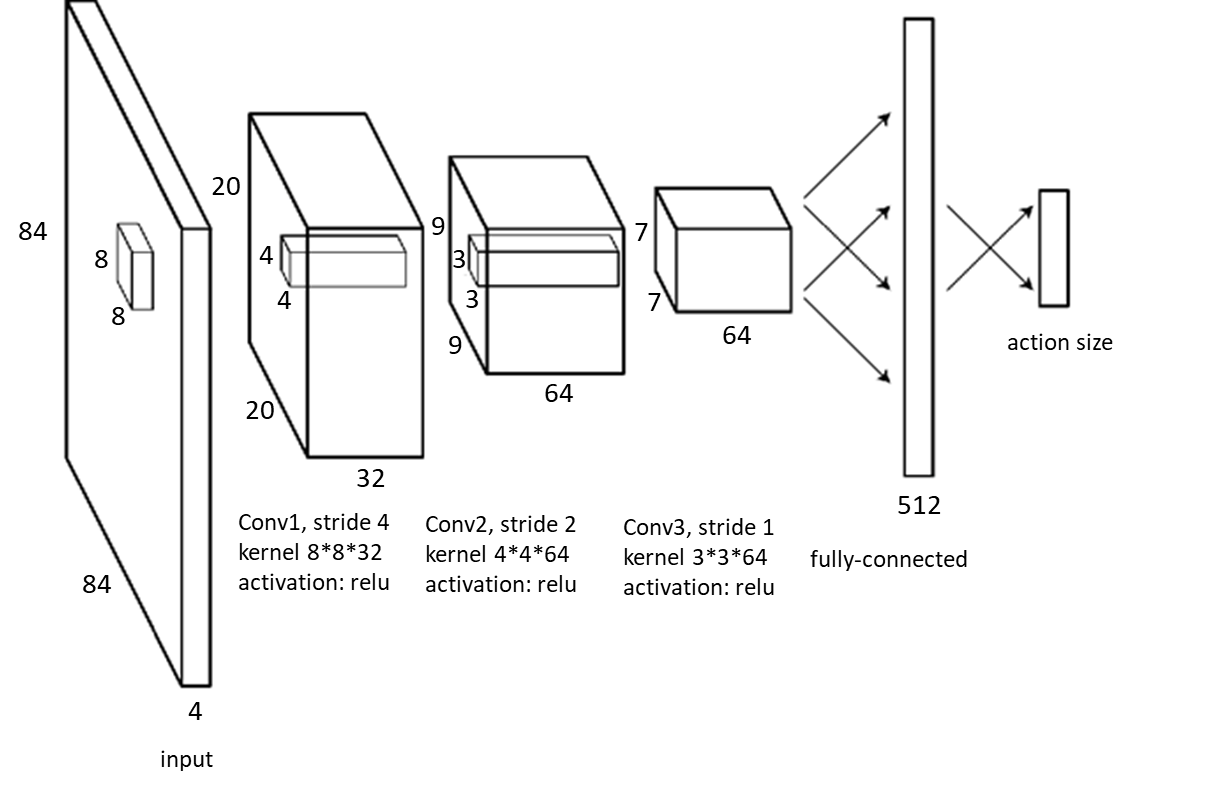
**4-2 Deep Q Learning**

1. Describe your DQN model

我們的Q-Network是由三層的convolutional layer與一層的fully-connected layer所組成，其詳細架構如圖一所示。



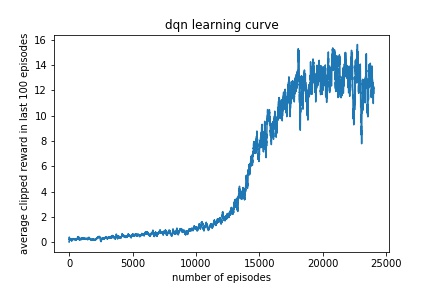
圖一、Q-Network之模型架構

Training details:

* Discount factor = 0.99
* Number of steps over which to anneal epsilon = 1000000
* Final value of epsilon in epsilon-greedy = 0.025
* Initial value of epsilon in epsilon-greedy = 1.0
* Number of previous transitions to remember = 10000
* Batch size = 32
* The frequency with which the target network is updated = 1000
* Number of episodes played = 25000
* Maximum number of steps played in an episode = 10000
* Optimizer = RMSPropOptimizer(learning\_rate=0.00015, decay=0.99)

1. Learning curve

* X-axis: number of episodes
* Y-axis: average clipped reward in last 100 episodes

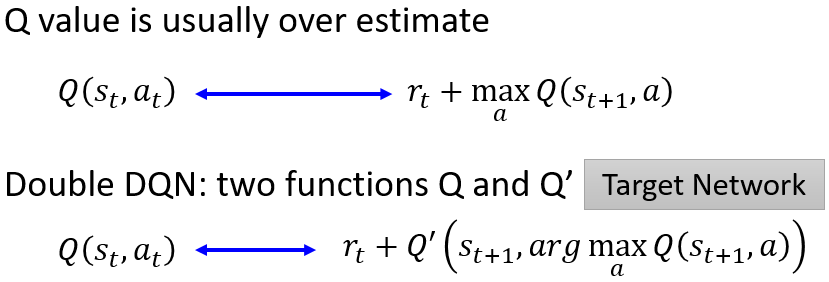


Implement 1 improvement method: Double DQN

1. Describe your tips for improvement

由於DQN很容易會over-estimate action的Q value，而且估計誤差會隨action的個數增加而增加。高估Q value的結果有可能會導致某個次好action的Q value超過了最好action的Q value，使得永遠無法找到最好的策略。

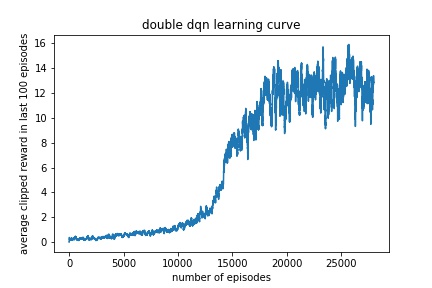
在Double DQN中，會先在Q Network找到 max *Q*(st+1,a) 的action，再去Target Network中找到這個action的Q value當成Target的Q value。這個Q value在Target Network中不一定是最大的，因此可以避免選到高估的次好Action。



圖二、Objective function of DQN and Double DQN

1. Learning curve

* X-axis: number of episodes
* Y-axis: average clipped reward in last 100 episodes



1. Compare to origin Deep Q Learning

從結果來看，DQN與Double DQN在訓練完成時都可以得到約70的reward，而Double DQN的結果並不會表現得特別好。另一方面，DQN與Double DQN幾乎都在第20000個episode時開始收斂。其中，Double DQN震盪的情形較DQN來的平緩。