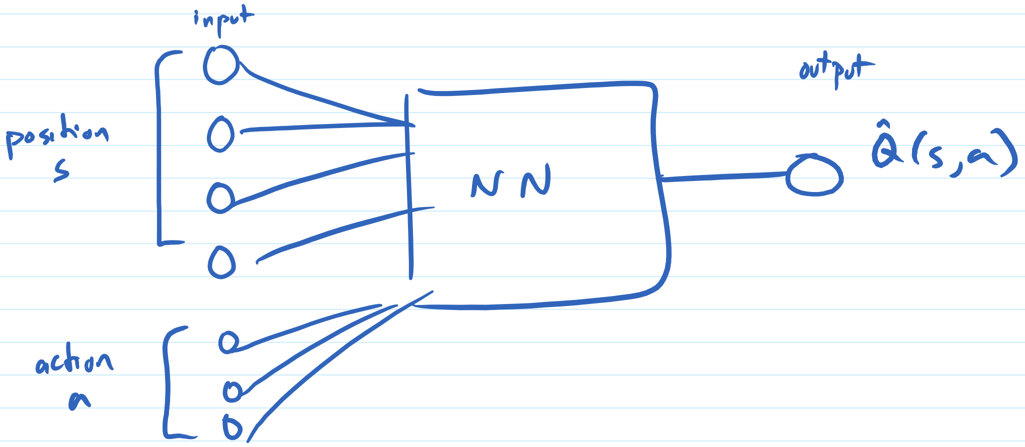
Deep Q Learning

```
initialize learning, target networks

for each iteration
  → for each episode          some smallish number
        for each event
            add (s, a, s', r) to replay database
    sample replay database
    train learning network towards $r + \gamma \cdot \underbrace{\max_{a'} \hat{Q}(s', a')}_{\text{computed by target network}}$

    if enough time has passed
        copy learning network to target network
```

ANN for approximating $Q(s,a)$



$Q((80,4,10,\cancel{24),\ 1)}) = 0.53\ldots$ $\hat{Q}((80,4,10,74),1) = 0.30$

observe $r + \gamma \underbrace{V(s')}$, $\rightarrow$ $\boxed{\max_{a} \hat{Q}(s', a)}$

and record in replay database (experience replay)

use as training input

expected $\hat{Q}(s,a)$

output of neural network

use 2 separate networks

1) to estimate future reward

"target" network

2) to train

learning or "prediction" network