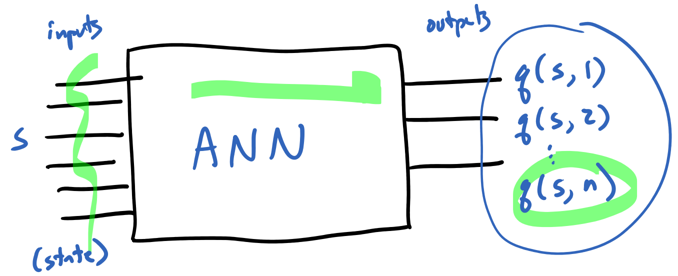
Deep Q Learning



$q(s,1) = .32$

$q(s,2) = .47$

$q(s,3) = .16$

choose $a=2$

s' has value $\underline{.46}$

$r = \underline{0.02}$

$q(s,2)$ updated toward $\underline{0.02 + \gamma \cdot 0.46}$

in training example

$\underline{X}$ s

$\underline{Y}$ $[q(s,1), 0.02+\gamma\cdot 0.46, q(s,3)]$

train this one

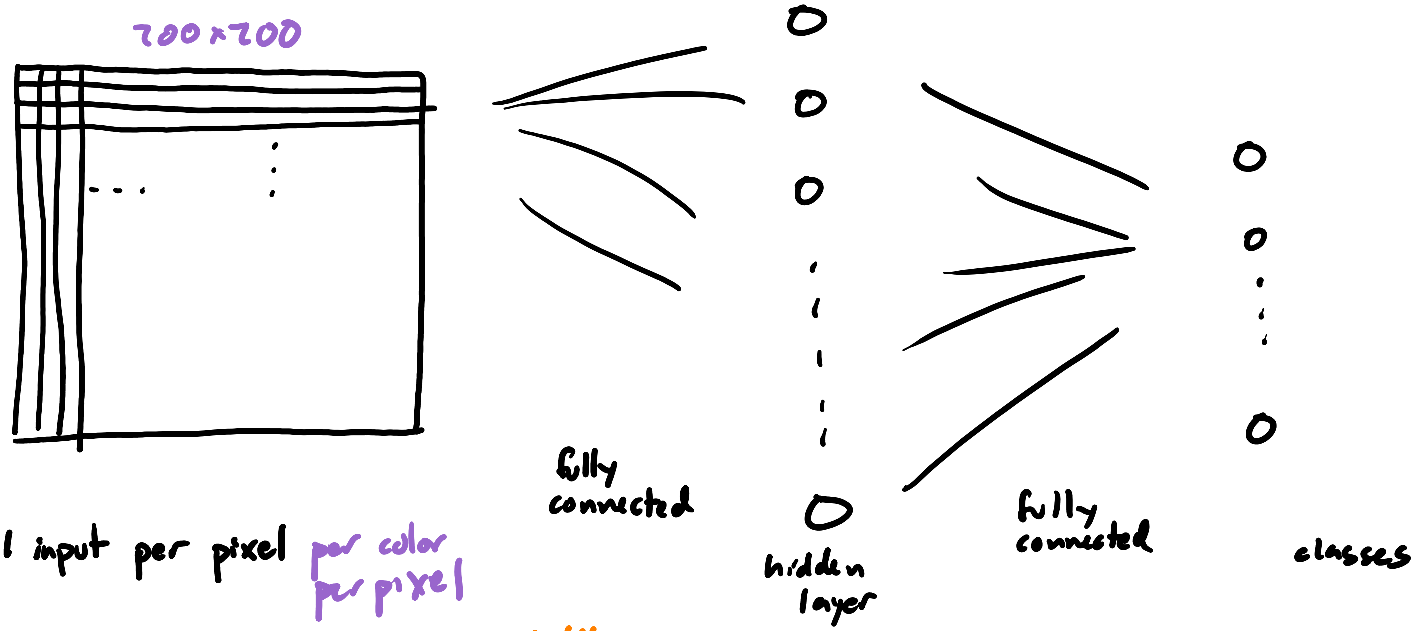
produces target output (what to train toward)

$\underline{q(s,a)} \rightarrow s', r$

$r + \gamma \underbrace{\max_{a'} q(s',a')}_{v(s')}$

```
initialize learning, target networks
for each iteration
    for each of n episodes          start each at initial state
        for each event
            add (s, a, s', r) to replay database
    sample replay database
    train learning network → target output [$\hat{q}_{learn}(s,1), \ldots, r+\gamma \max_{a'} \hat{q}_{target}(s,a), \ldots, \hat{q}_{learn}(s,n)$]
    if enough time passed
        copy learning network to target network
```

ANNs for Images



120k inputs

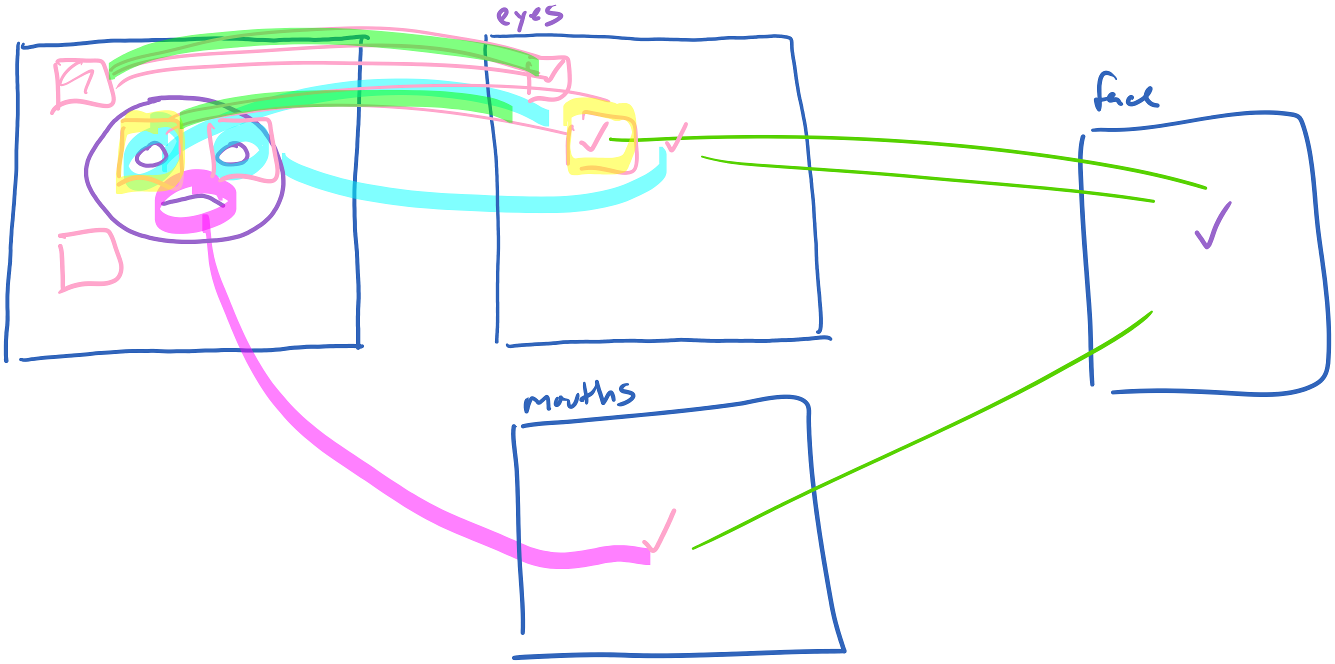
14400m

120k nodes

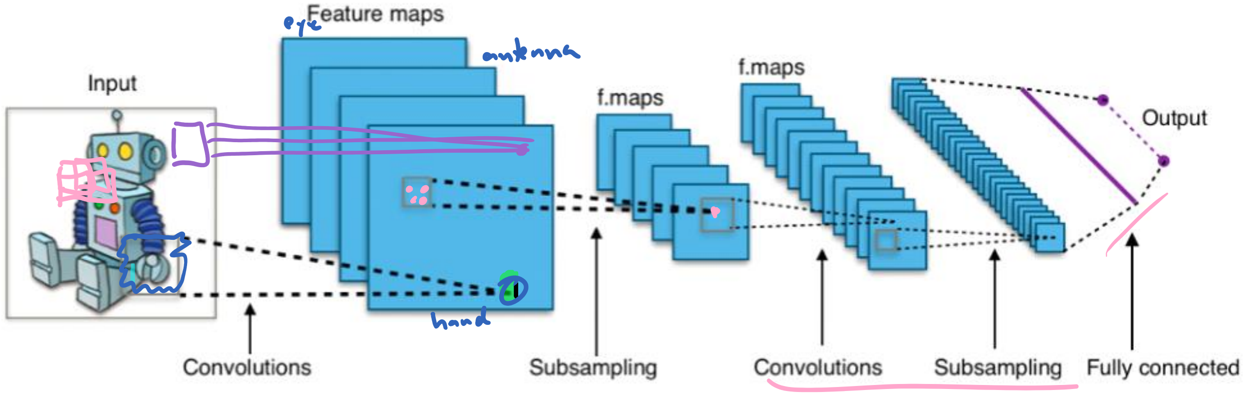
14B weights

Convolutional Neural Networks → NNs for image processing

Deep Q network learning to play Pong



A much better picture from wikipedia user aphex34, who does not endorse these notes.



https://upload.wikimedia.org/wikipedia/commons/6/63/Typical_cnn.png

AlphaGo (2014-2017)

Step 1: supervised learning for convolutional deep neural network (policy network)

DB of expert-level games

13 layers

input: 19×19×48

black
white
empty
#opp captured
#own captured
liberties
ladder capture
ladder esc
⋮

output: action (19×19+1 outputs)

3 weeks

~ matched 55% of time

+ smaller (faster) network 25% of time

Step 2: reinforcement learning for convolutional deep neural network

1 day

beat SL 80% of time

Step 3: reinforcement learning for value network

output is v(s)

using step 2 network
plays itself 30M times
sample 1 pos/game

Step 4: MCTS

default: use fast network from step 1

initialize new node's value using step 3 value network

tree policy: $\underbrace{q(s,a)}_{\text{exploit observations}} + c\,\underbrace{P(s,a)}_{\text{from larger step1 network}} \cdot \frac{\sqrt{\#\text{parents visited}}}{1+\#\text{child visits}}$

Elo 3144 → 3739 → 5185

2015 (Fan Hui) 2016 (Lee Sedol) 2017 (retired)

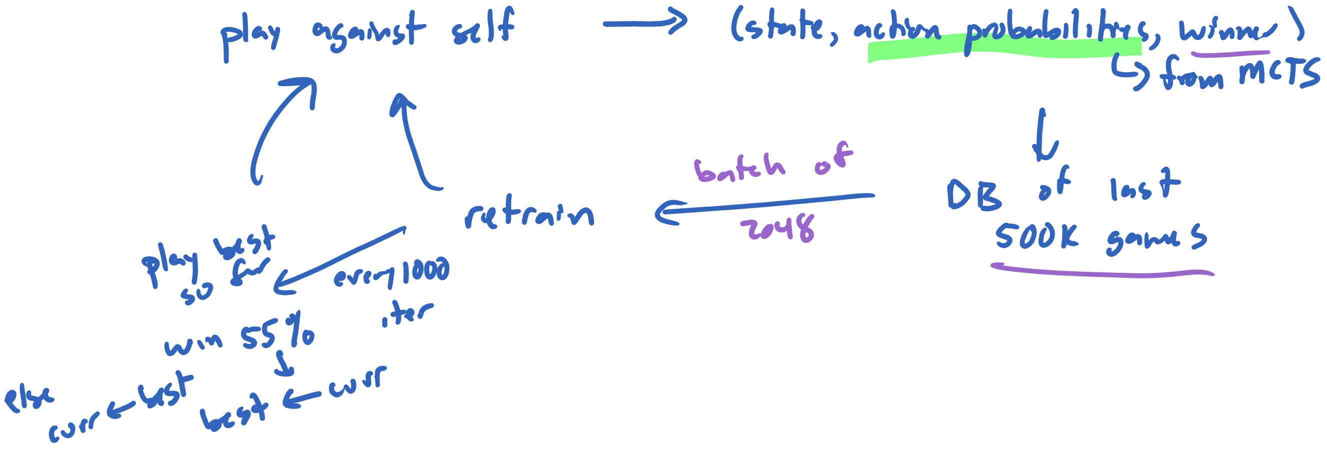
Δ Elo 400 → higher rated player has 90%+ chance of winning

AlphaGo Zero

no prior knowledge

input : 19 x 19 x 17   current pos + last 7 pos + turn (all 1= black 0= white)

output : move (19x19+1) and valu [-1, +1]



[Deep Reinforcement Learning Doesn't Work Yet (alexirpan.com)](https://alexirpan.com)