Playing Games
The Early Days: Chess

Exhaustive tree search with pruning (alpha-beta)

Handwritten evaluation: $V(s)$
Deep Learning: Atari Games

Preprocessing
Conv 8x8, C=16, stride=4
ReLU
Conv 4x4, C=32, stride=2
ReLU
Linear (256)
ReLU
Linear (4-18)

"Deep Q learning" Learn to predict the long-term rewards for each state-action pair

\[ V(s) = \mathbb{E}_{a \sim \pi(s)} Q(s, a) \]

Superhuman performance on three of the seven games they tested on.

Deep Learning + Search: Go

• Two network outputs: policy and value
  \[ \pi, V \]

• Monte Carlo Tree Search


1: AlphaGo
2: AlphaGo Zero

Sample rollouts
\[ a_i \sim \pi(s_i) \]

Evaluate the resulting position
\[ V(s_N) \]

AlphaGo: Initially trained by imitation learning
Later trained by self-play

AlphaGo Zero: Trained entirely by self-play
AlphaGoZero Architecture

Conv 3x3, C=256
BatchNorm
ReLU

Conv 3x3, C=256
BatchNorm
ReLU

Conv 3x3, C=256
BatchNorm
ReLU

Conv 1x1, C=2
BatchNorm
ReLU
Linear 362

Conv 1x1, C=1
BatchNorm
ReLU
Linear 256
Linear 1
Tanh

40 X

Action

Value
Generalization: AlphaZero and MuZero

- **AlphaZero**: Same architecture for Chess, Shogi, and Go.
- **MuZero**: No prior knowledge of the game rules.


Sample rollouts $a_i \sim \pi(s_i)$

Evaluate the resulting position $V(s_N)$

Rollout using a *learned* model of the environment

“Model-based reinforcement learning”
NNUE: Shogi and Chess

Deep networks are slow to run on CPU

Allows efficient computation of adjacent positions

Special Embedding
- Linear 256
- Linear 256
- Linear 32
- Linear 32
- Linear 1

Most activations are unchanged

~500x more nodes per second than Leela Chess Zero

Video Games

- Partial observability, complex rules
- High-dimensional state and action spaces
- Real-time constraints
- Cooperation
Dota 2: OpenAI 5

Trained with self-play and reward shaping.
Defeated top 1v1 players in August 2017.
Won 5v5 match against the world champions in April 2019.

OpenAI, et al. “Dota 2 with Large Scale Deep Reinforcement Learning.”
StarCraft II: AlphaStar

Combination of imitation learning and self-play

Rating higher than 99.8% of human players in blind league play

Feeback to Human Play

Ichiriki Ryo (9p) vs Byun Sangil (9p)
14th Chunlan Cup World Weiqi Championship Round of 24

Magnus Carlsen vs. Shakhriyar Mamedyarov
Norway Chess 2019 Round 4