CS 181V:
Reinforcement Learning

Lecture 25
April 29, 2020
Neil Rhodes

Images from:

- **David Foster, Applied Data Science**

- **https://medium.com/applied-data-science/how-to-build-your-own-muzero-in-pytho f77d5718061a**

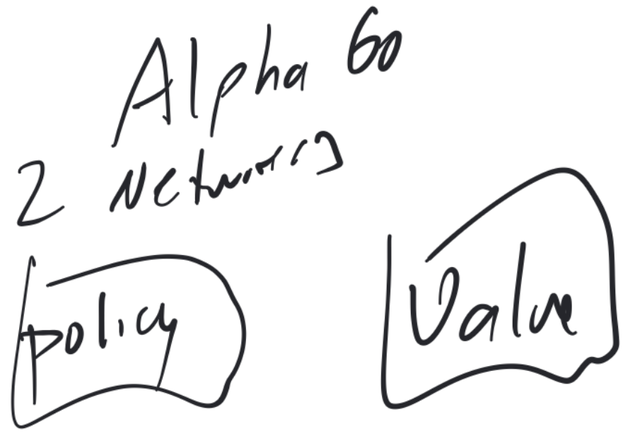
- ***Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm* by Silver, et al., 2017**

- ***Mastering the game of Go without human knowledge*, Silver, et al., 2017**

History

- AlphaGo, 2015
- Hand-crafted features and trained on human plays
- Two Neural Networks: One to learn the value function, one to learn policy function
- Policy function trained on large body of human plays

History



- AlphaGo Zero, 2017
  - No prior knowledge of Go (no hand-crafted features, no training on human games). All self-play
  - Single Residual Neural Network learns both value and policy functions
  - Keeps best-neural-network-so-far (new champion if beats >55% of games against old champion)

AlphaGo Zero

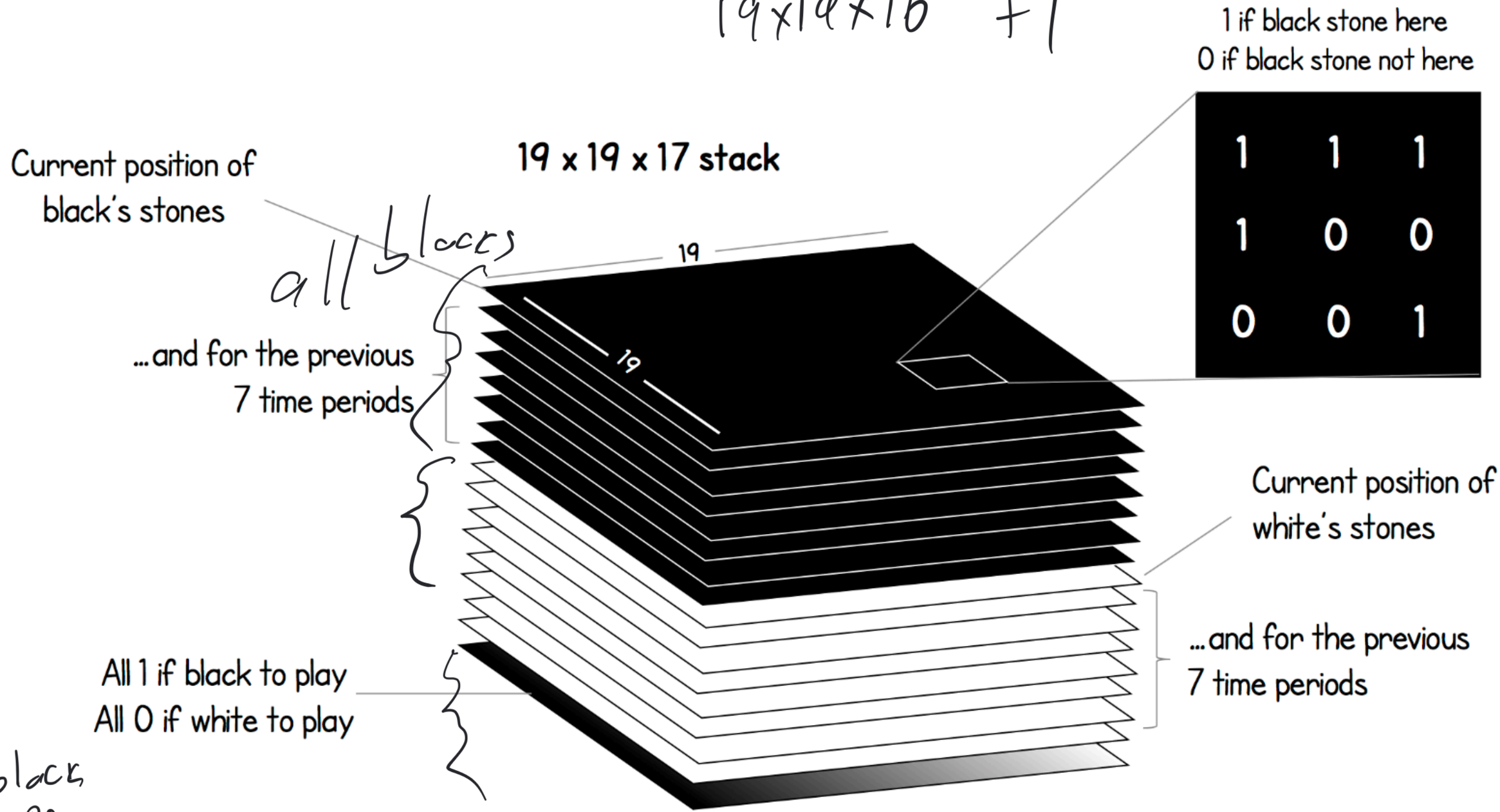
Learn through self-play

- Get rid of handcrafted features
- No network weights pretrained from human games
- One shared network to compute both policy and value
- Simplified MCTS: no rollouts!

AlphaGo Zero

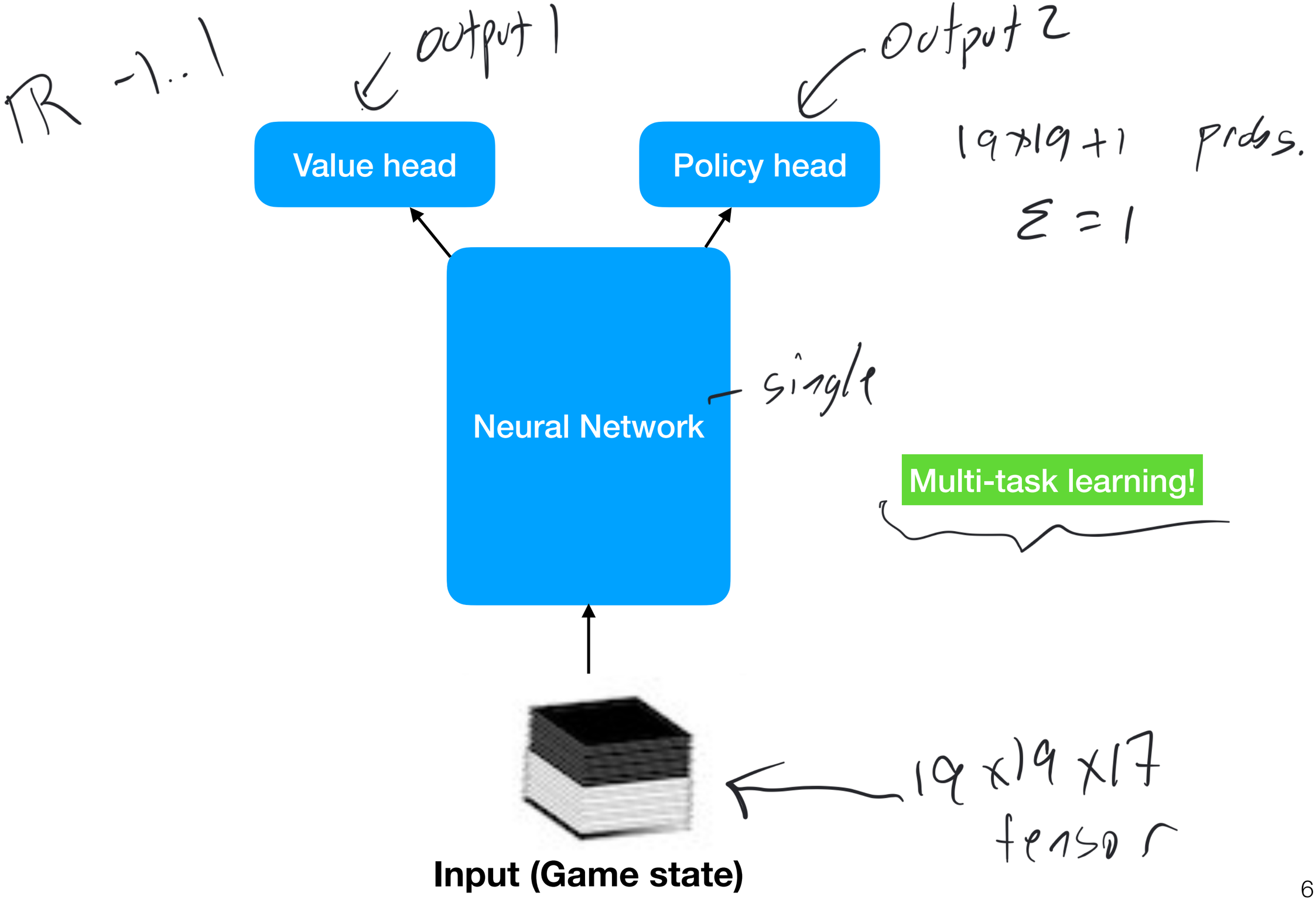
AlphaGo Zero: What is a game state for Go?

19x19x16 +1



AlphaGo Zero

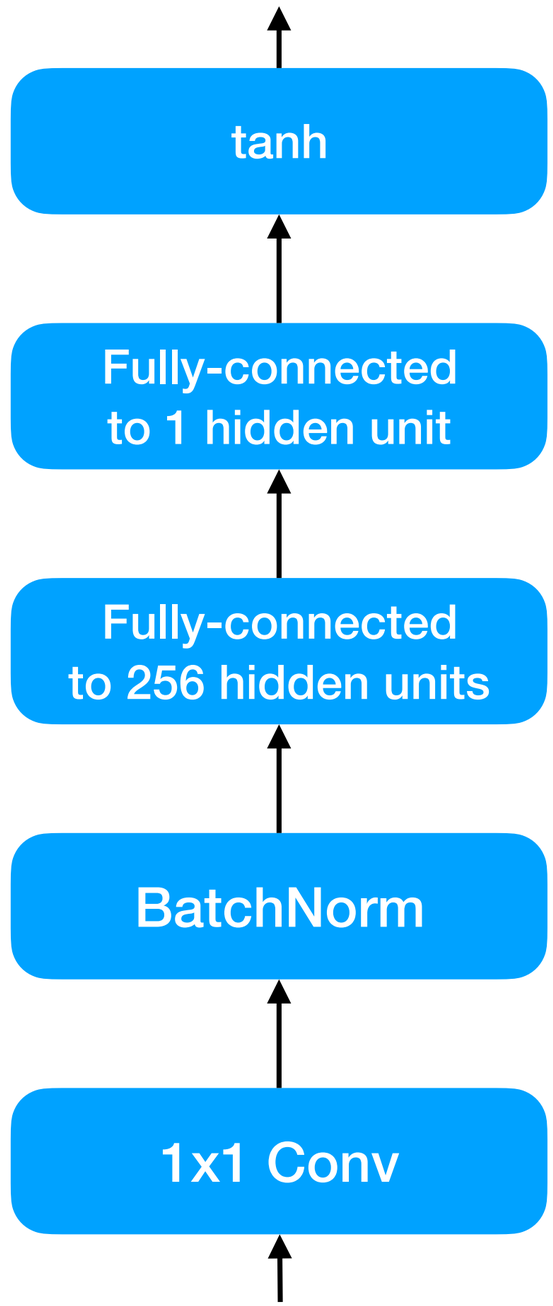
The Neural Network



AlphaGo Zero

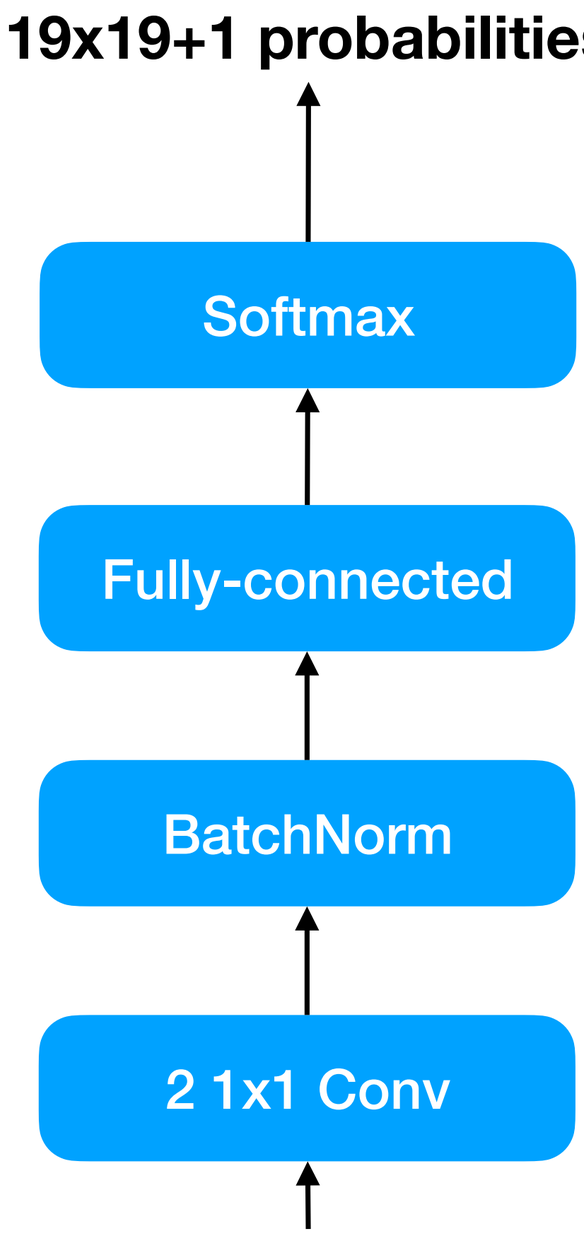
Value (V) Head

$\mathbb{R}\in$[-1, 1]: game value for current player (-1=lose, 0=draw, +1=win)



AlphaGo Zero

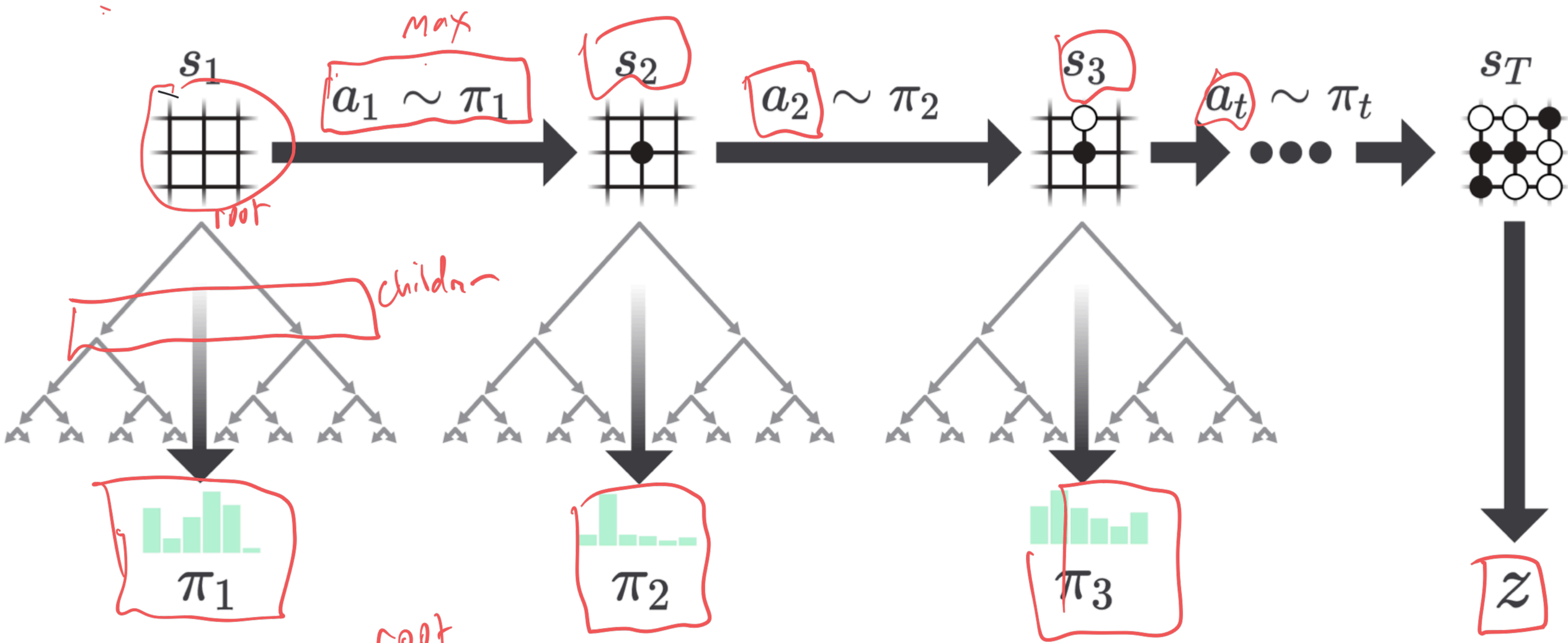
Policy (P) Head



19x19 places to play stone
1 way to pass

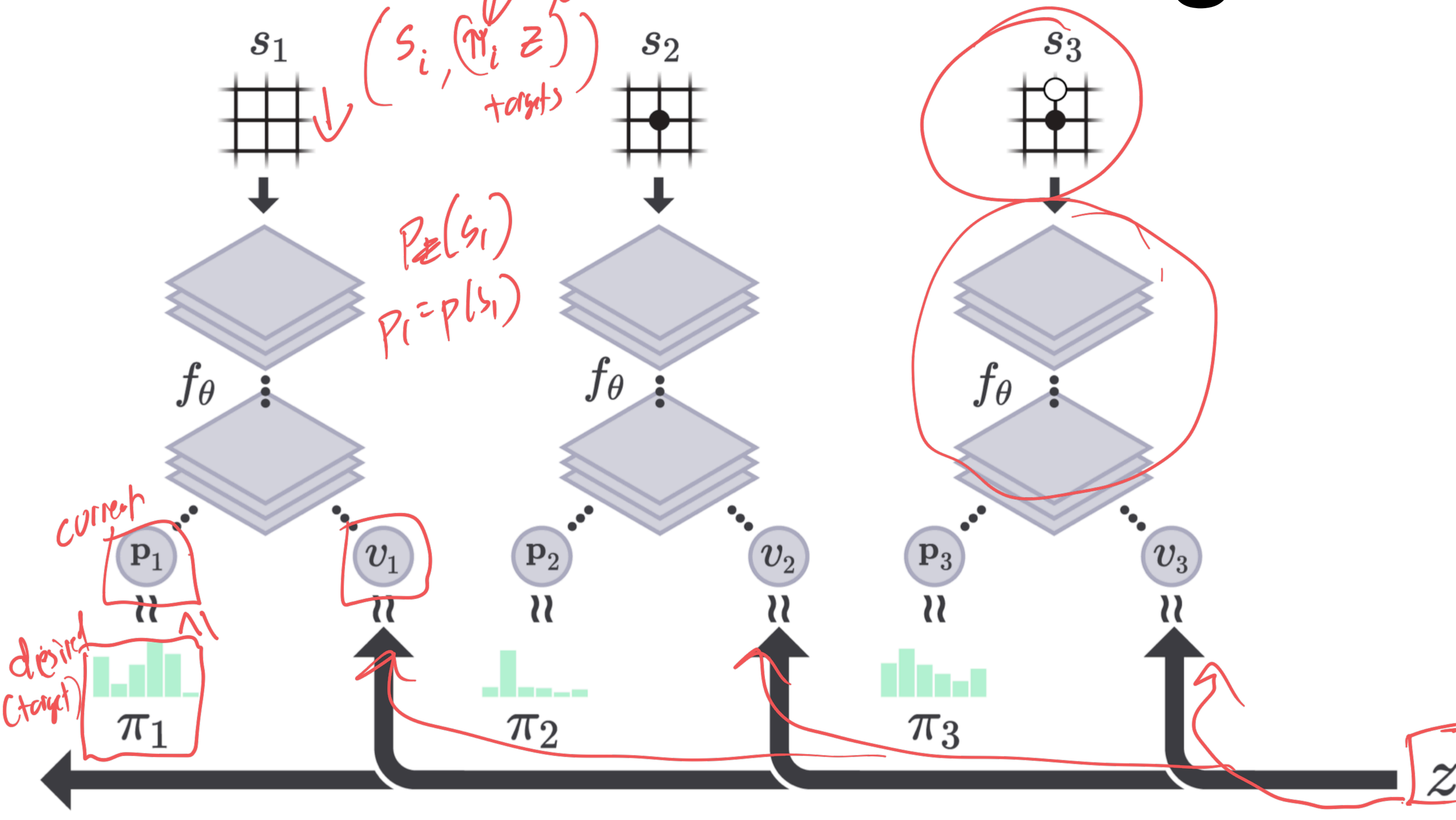
AlphaGo Zero

Self-play



AlphaGo Zero

Neural Net Training



$$L = (z - v)^2 - \pi^T \log p + c\|\Theta\|^2$$

AlphaGo Zero

Neural Net Training

- Update Θ so that:
  - Given an input state, s_i, f_Θ produces output that is closer to:
    - For value portion: z_i the actual result of the game
    - For action portion: π_i the MCTS-improved policy based on the neural net's p_i

$f_\Theta(s_i) \rightarrow p_i, v_i$

AlphaGo Zero

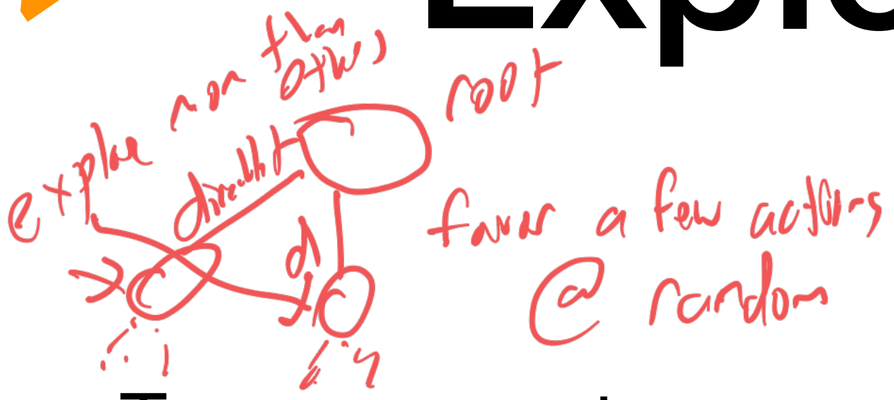
Monte Carlo Tree Search (MCTS)

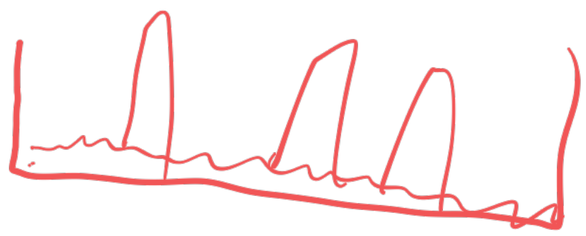
Each state has edges for all legal actions:
For each edge keep:
N(s, a): **how many times has this state/action pair been seen**
W(s, a)**: total action-value**
Q(s, a)**: mean action-value: *W(s, a)/N(s, a)***
P(s, a)**: prior probability of selecting that edge**

from NN

AlphaGo Zero

Explore/Exploit



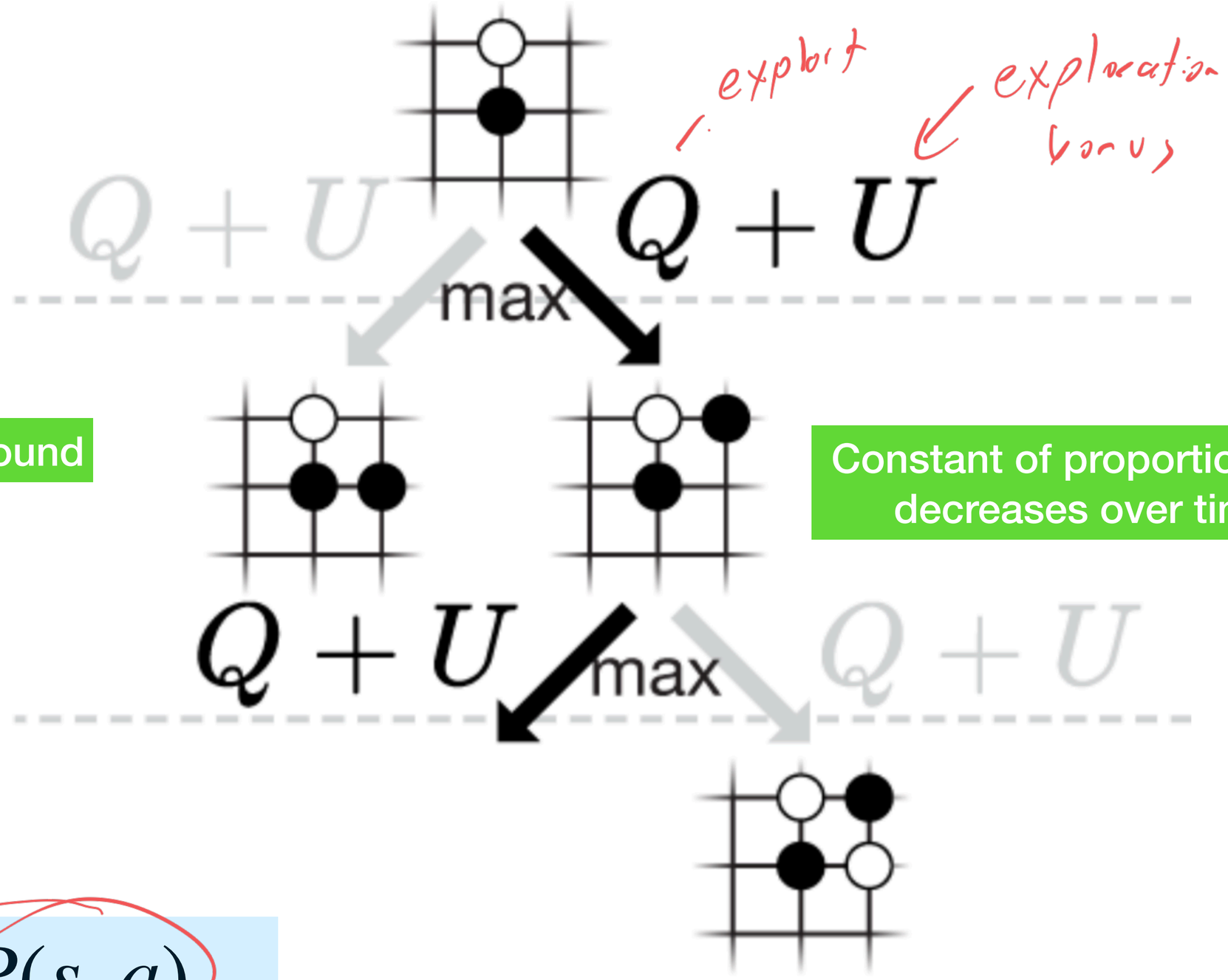


- Two sources to encourage exploration:
  - Dirichlet noise added to top-level P(s) of MCTS
    - Non-zero chance of any move happening
  - Upper-confidence bound when evaluating move in MCTS
    - Encourage actions whose confidence is low

Dirichlet noise: sums to given value (so can be a probability), and, with parameter used in AlphaGoZero,
makes most of the probability focused in small area

AlphaGo Zero

MCTS: Select



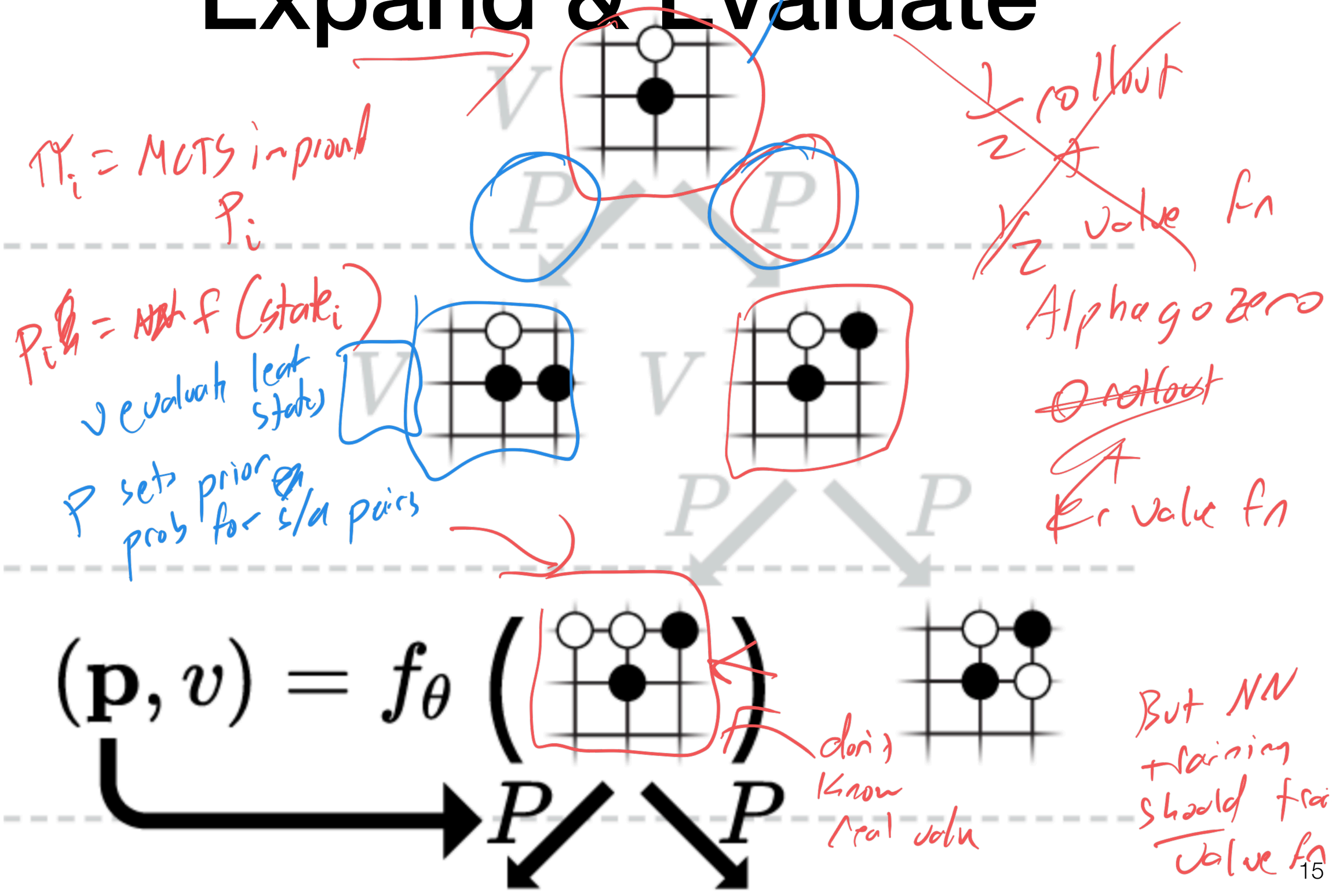
U Is upper-confidence bound

Constant of proportionality decreases over time

$$U(s,a) \propto \frac{P(s,a)}{1 + N(s,a)}$$

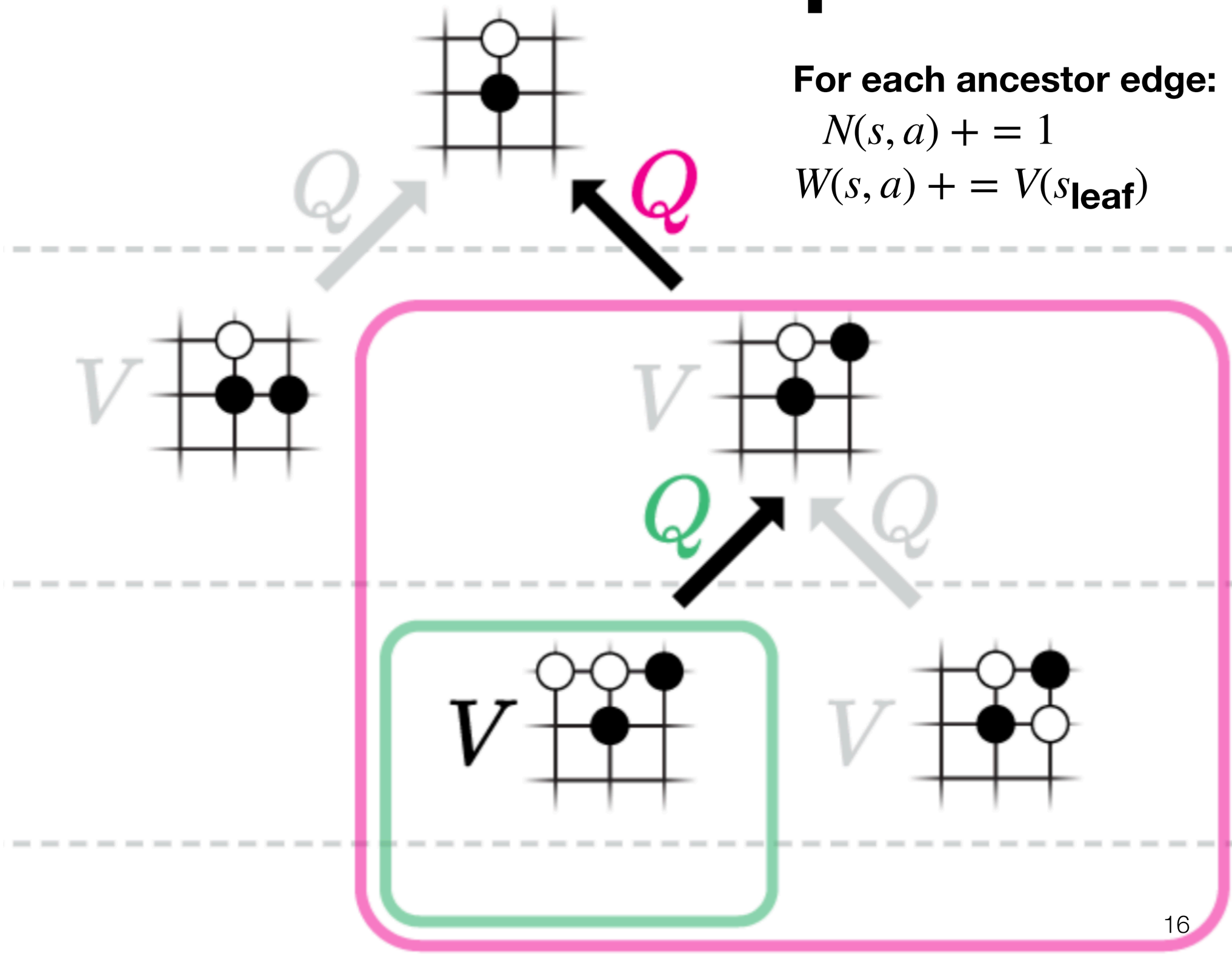
AlphaGo Zero

MCTS: Expand & Evaluate



AlphaGo Zero

MCTS: Backup



For each ancestor edge:

$$N(s,a) += 1$$

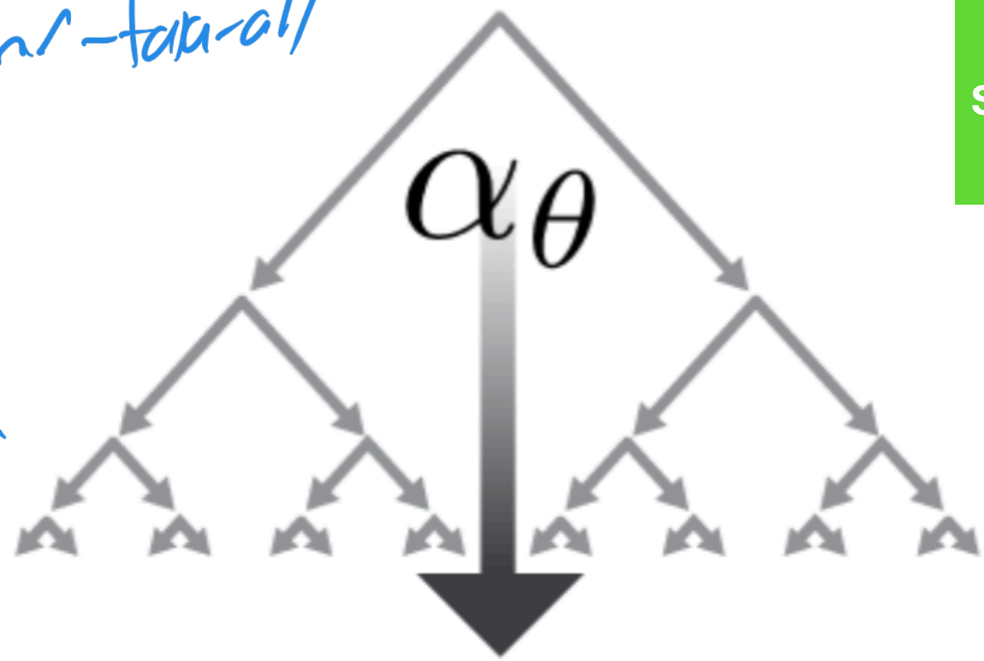
$$W(s,a) += V(s_{\textbf{leaf}})$$

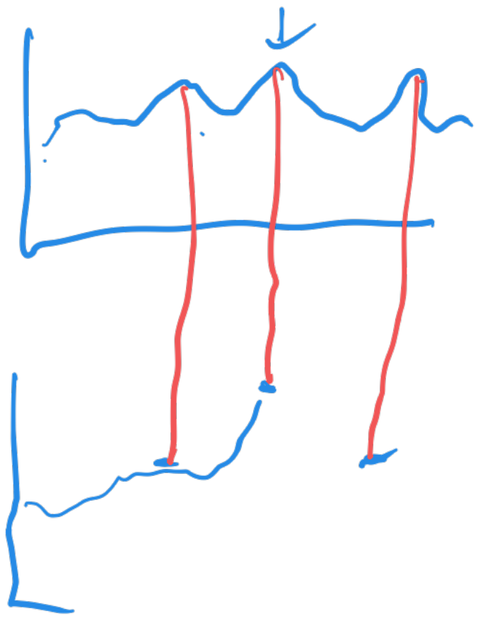
AlphaGo Zero

MCTS: Play

$$\pi(a \mid s_0) = \frac{N(s_0, a)^{1/\tau}}{\sum_b N(s_0, b)^{1/\tau}}$$

lower temperature causes it to become winner-take-all





½ power

¼ power more accentuated

τ is a temperature constant: starts near 1 and ends close to zero. For play: close to zero

@ beginning we're not wedded to our winner

After an action is chosen from π, keep tree starting from resulting state, clear rest of tree

AlphaGo Zero

What Go-specific knowledge is used?

- Network knows input format (19x19 board, etc.)
- Network knows how many possible actions (19x19+1)
- MCTS knows which actions lead to which states
- MCTS knows whether a state is terminal
- MCTS knows how to score a terminal state
- MCTS does dihedral reflection or rotation (takes advantage of symmetry of Go board)

AlphaZero

Remove the Go

History

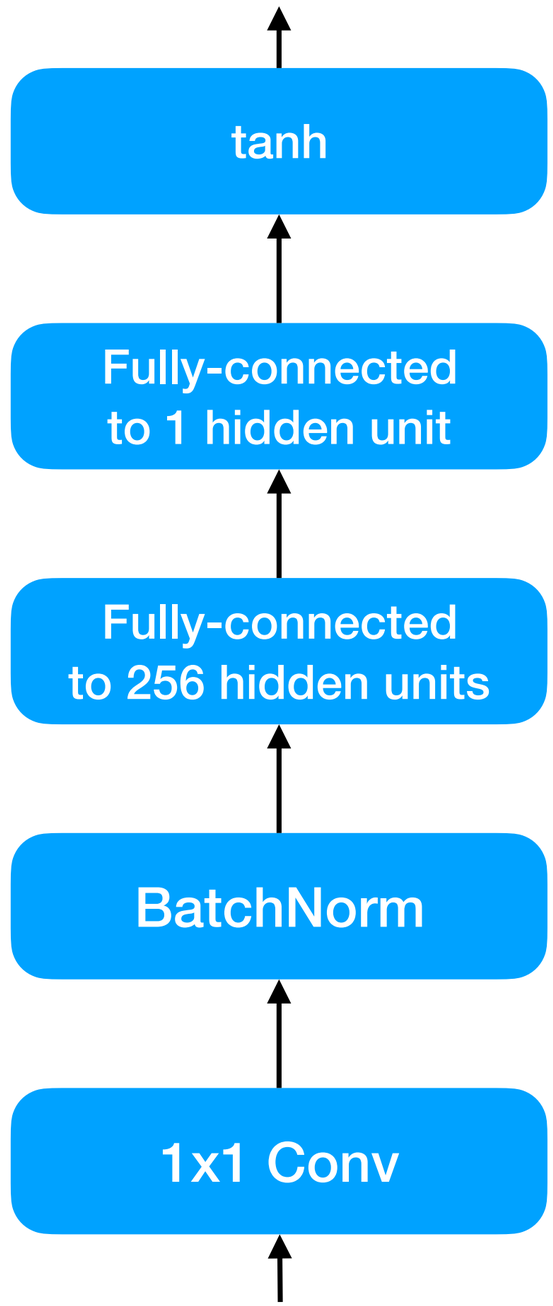
- AlphaZero, 2017
  - No current best Neural Net. Always uses the latest NN.
  - Same network (other than input and head) with almost same hyper-parameters can learn Go, Chess, and Shogi
    - One hyper-parameter is scaled based on number of possible actions

AlphaZero

Value (V) Head

independent of the game

$\mathbb{R} \in$ [-1, 1]: game value for current player (-1=lose, 0=draw, +1=win)



AlphaZero

Domain knowledge

- Network knows input format
- Network knows how many possible actions
- MCTS knows which actions lead to which states
- MCTS knows whether a state is terminal
- MCTS knows how to score a terminal state
- Typical number of legal moves scales exploration noise

different rules for diff games

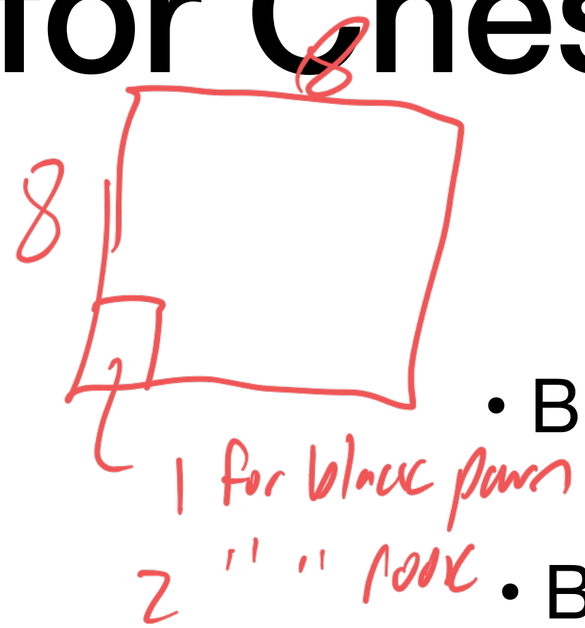
dirichlet

AlphaZero

What is a game state for Chess?

8x8x119 stack

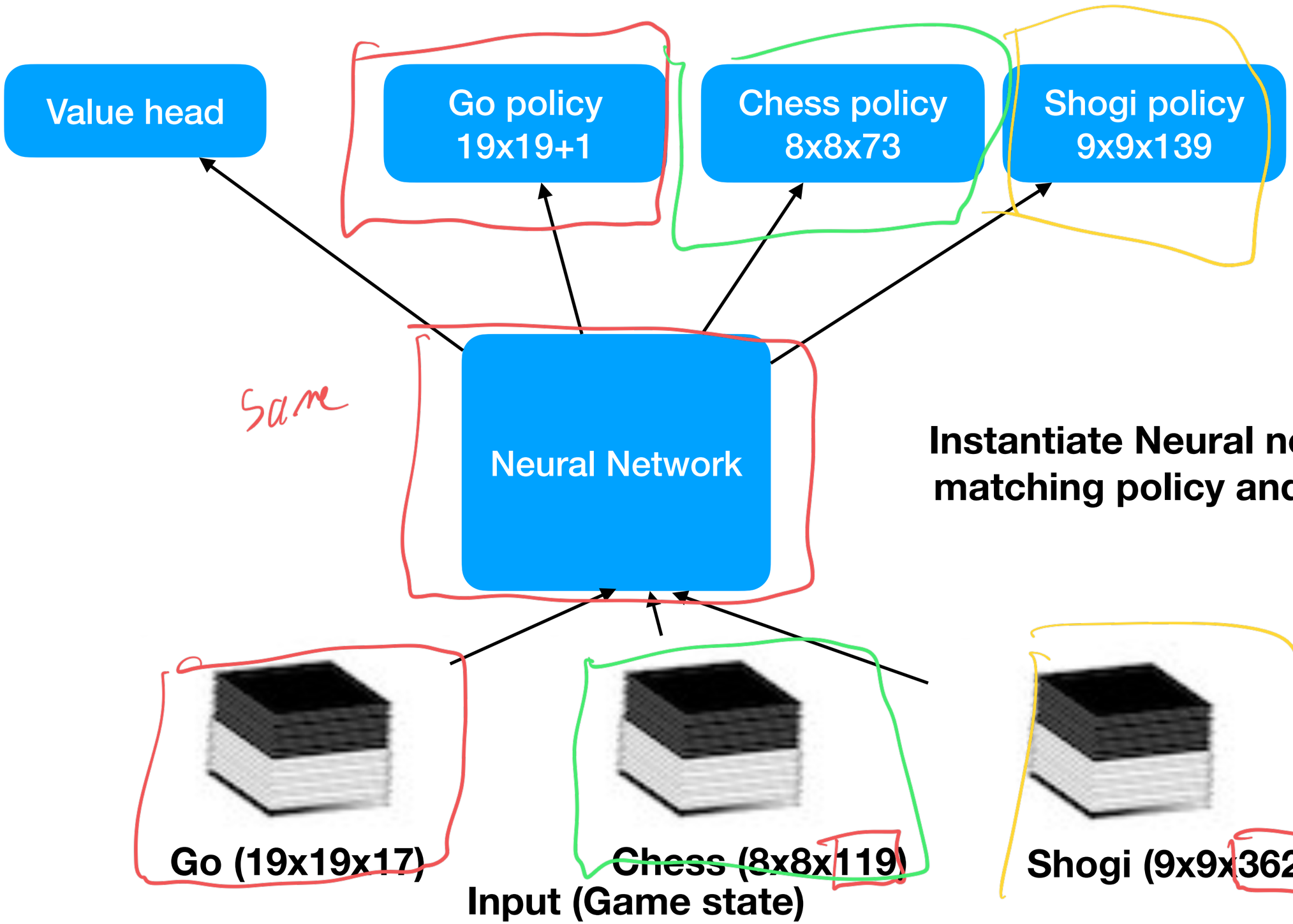
- 8x8 plane:
  - White pawns
  - White Rooks
  - White Bishops
  - White Knights
  - White Queens
  - White King
  - Black pawns
  - Black Rooks
  - Black Bishops
  - Black Knights
  - Black Queens
  - Black King



+ Extra planes for: repetition, color, total move count, and a few others

AlphaZero

The Neural Network

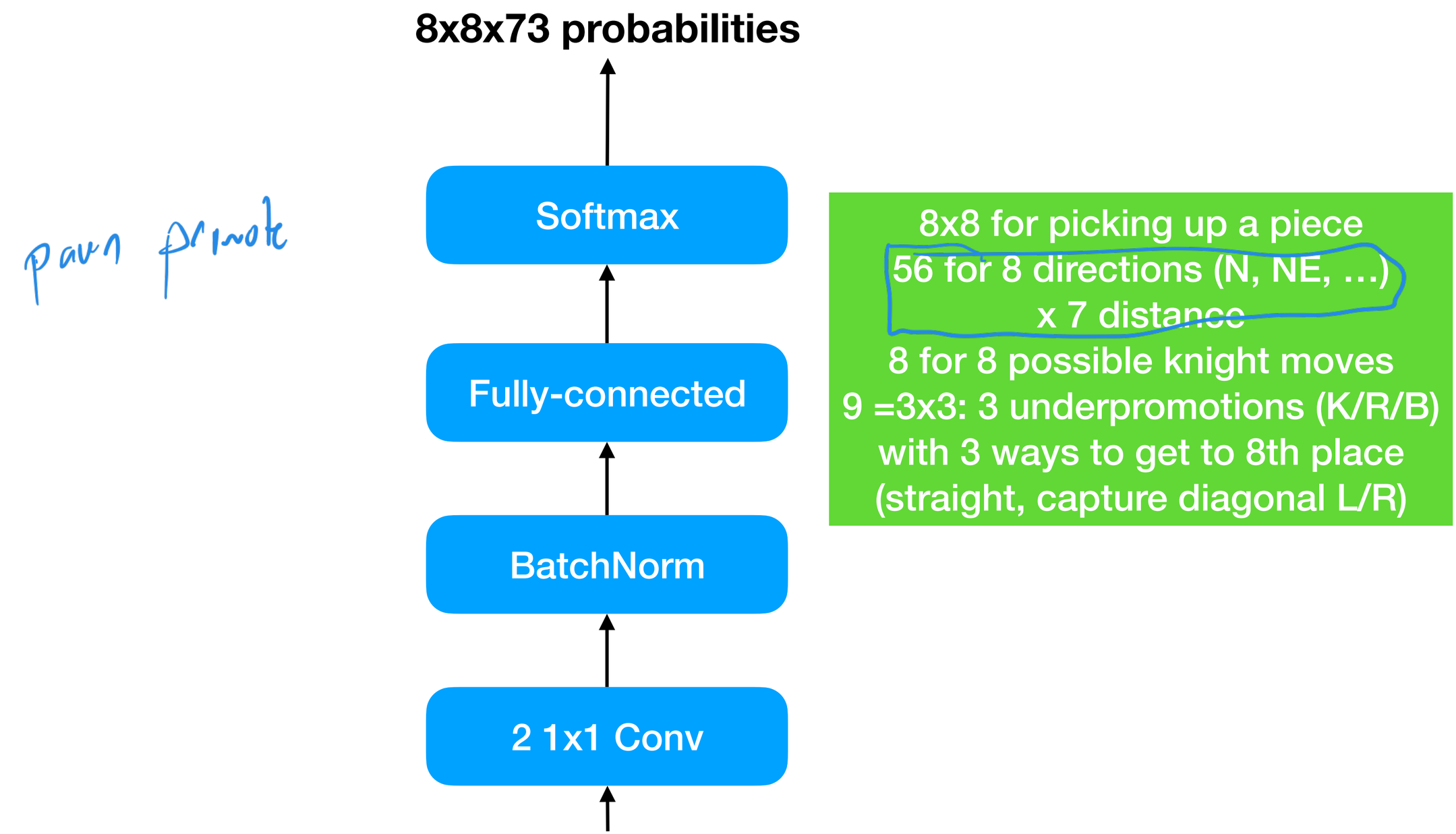


Instantiate Neural net with matching policy and input

AlphaZero

Policy (P) Head (for Chess)

From : 8×8



pawn promote

AlphaZero

Generalized approach & NN itself

Results

Works great

less filling

- Superhuman results on Go, Chess, Shogi
- Beat existing computer players

AlphaZero

Takeaways

- Simple representation of input and output
- MCTS improves P to π

Alpha Go & Alpha Go Zero

- Train NN:x= game state, y=(π, game_result), ŷ=(P, V)

input

target

grand truth value

policy

Value

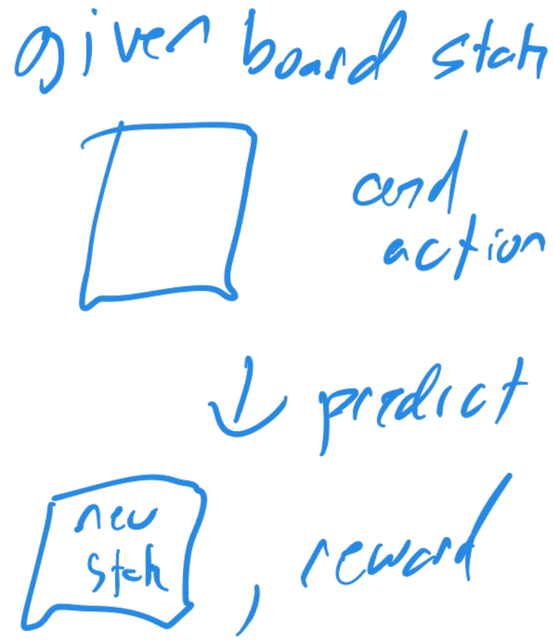
MuZero

History

- MuZero, 2019
  - Doesn't know the rules of the game (During MCTS, must use learned representation of the game dynamics)
  - Is told what moves are legal in the current position
  - Is told when the game is over (or it's a draw) and who won
- Extended to work with Atari games as well as Go:
  - Is also told points earned at each step

MuZero

Planning with a Learned Model



- Network knows input format
- Network knows how many possible actions
- ~~MCTS knows which actions lead to which states~~
- ~~MCTS knows whether a state is terminal~~
- ~~MCTS knows how to score a terminal state~~
- Typical number of legal moves scales exploration noise

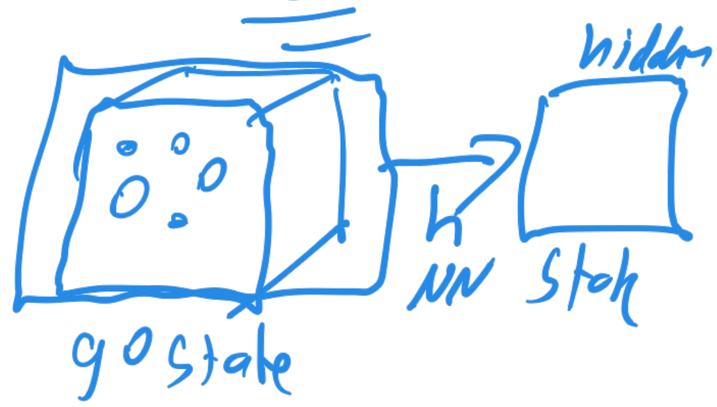
MuZero

How to do MCTS without knowing the dynamics?

- Solution: Train a network to generate the dynamics
- Key insight: Don't work with *real* states (o_k), instead, use *hidden* state representations (s_k)

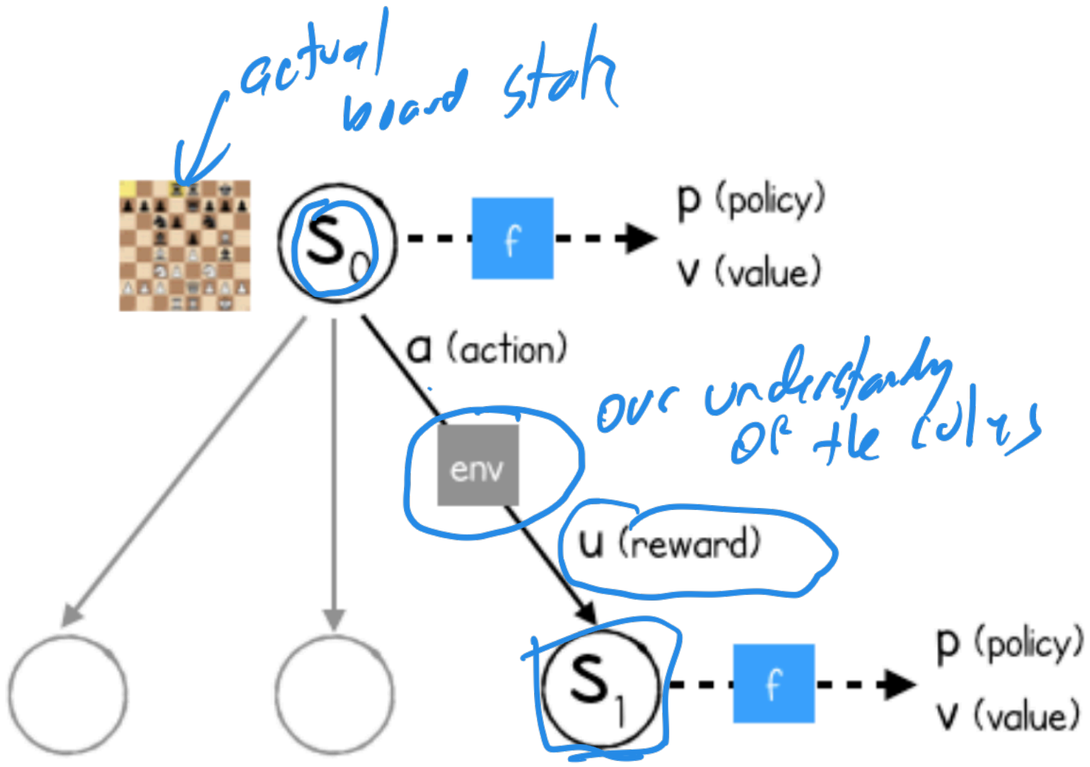
o_k = observation k
real state

- Train jointly three functions (NN):
  - representation h_θ: $o_k \rightarrow s_k$
  - dynamics g_θ : $s_k, a_k \rightarrow r_{k+1}, s_{k+1}$
  - prediction f_θ: $s_k \rightarrow p_k, v_k$



MuZero

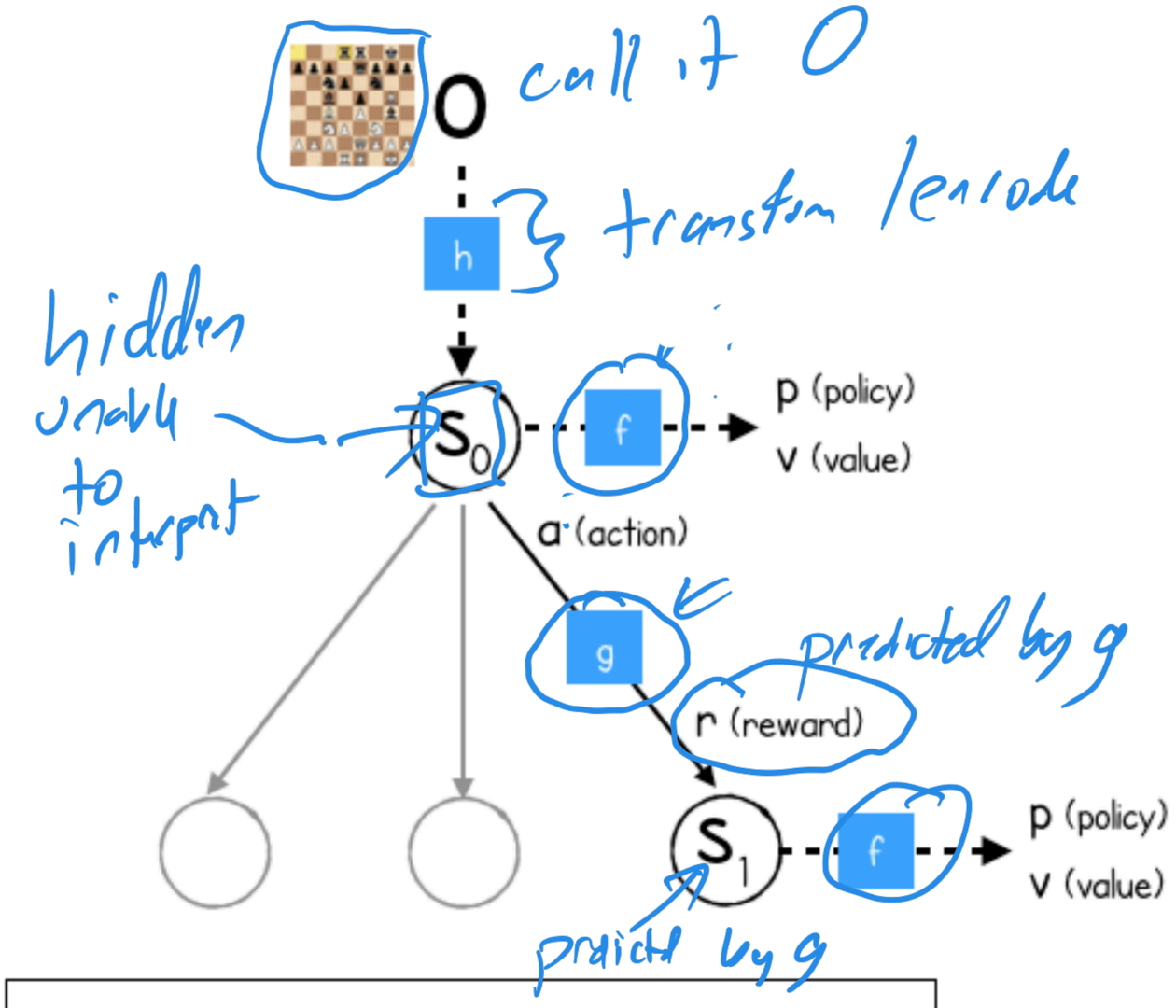
AlphaZero



AlphaZero has 1 network

		from	to
prediction	f:	s	→ p, v

MuZero

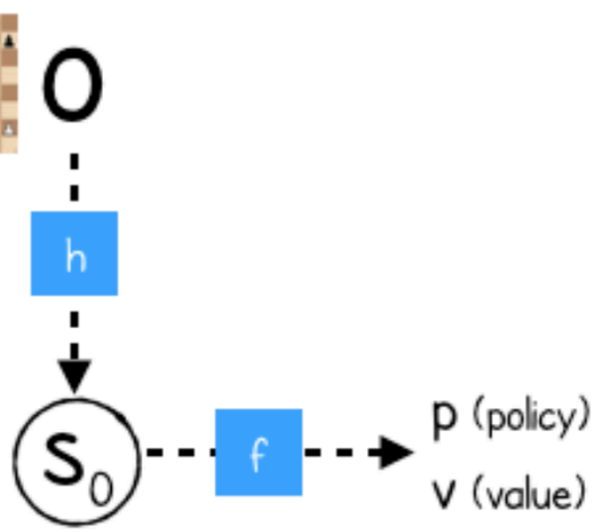


MuZero has 3 networks

		from	to
prediction	f:	s	→ p, v
dynamics	g:	s, a	→ r, s
representation	h:	o	→ s

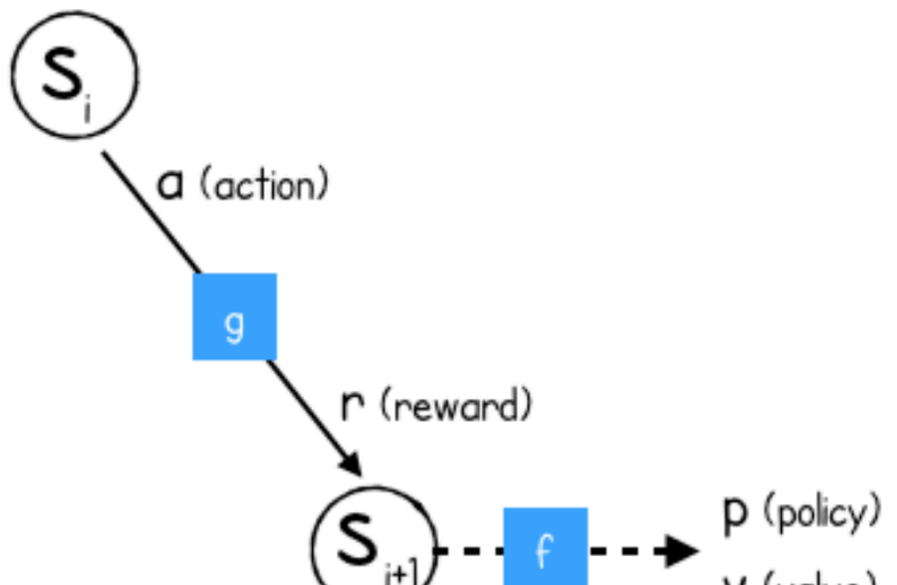
MuZero

initial_inference



input		output	
image	o	value	v
		reward	r (=0)
		policy_logits	p
		hidden_state	s_0

recurrent_inference



input		output	
hidden_state	s_i	value	v
action	a	reward	r
		policy_logits	p
		hidden_state	s_{i+1}

MuZero

MCTS:

- Inputs:
  - current real state (o_i)
  - functions $f_\theta, g_\theta, h_\theta$
- Outputs:
  - Improved policy for state o_i: π_i
  - Improved value for state o_i: ν_i

MuZero

Training

- We train f, g, and h to:
  - **Predict v and p well** (given hidden state)
  - **Predict reward well** (and generate useful new hidden state)
  - Predict initial hidden state well
- Not just for one timestep, but for several timesteps in the future

MuZero

Game 5, observation 42

	Time	Saved					Predicted/Calculated		
		State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
Unroll steps	42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}			
	43		u_{43}	a_{43}	π_{43}	ν_{43}			
	44		u_{44}	a_{44}	π_{44}	ν_{44}			
	45		u_{45}	a_{45}	π_{45}	ν_{45}			
	46		u_{46}	a_{46}	π_{46}	ν_{46}			

MuZero

Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}			
43		u_{43}	a_{43}	π_{43}	ν_{43}			
44		u_{44}	a_{44}	π_{44}	ν_{44}			
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

MuZero

Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}			
43		u_{43}	a_{43}	π_{43}	ν_{43}			
44		u_{44}	a_{44}	π_{44}	ν_{44}			
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

0 steps ahead

$$s_{42}^{0} = h_\theta(o_{42})$$

MuZero

Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}			
44		u_{44}	a_{44}	π_{44}	ν_{44}			
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

$$s^0_{42} = h_\theta(o_{42})$$

$$v^0_{42}, p^0_{42} = f_\theta(s^0_{42})$$

MuZero Predict several steps ahead

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}		
44		u_{44}	a_{44}	π_{44}	ν_{44}			
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

$$s^0_{42} = h_\theta(o_{42})$$

$$r^1_{42}, s^1_{42} = g_\theta(s^0_{42}, a_{42})$$

$$v^0_{42}, p^0_{42} = f_\theta(s^0_{42})$$

MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}			
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

$$s^0_{42} = h_\theta(o_{42})$$
$$r^1_{42}, s^1_{42} = g_\theta(s^0_{42}, a_{42})$$
$$v^0_{42}, p^0_{42} = f_\theta(s^0_{42})$$
$$v^1_{42}, p^1_{42} = f_\theta(s^1_{42})$$

MuZero Predict several steps ahead

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}		
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

$$s^0_{42} = h_\theta(o_{42})$$
$$r^1_{42}, s^1_{42} = g_\theta(s^0_{42}, a_{42})$$
$$r^2_{42}, s^2_{42} = g_\theta(s^1_{42}, a_{42})$$

$$v^0_{42}, p^0_{42} = f_\theta(s^0_{42})$$
$$v^1_{42}, p^1_{42} = f_\theta(s^1_{42})$$

MuZero

Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}			
46		u_{46}	a_{46}	π_{46}	ν_{46}			

$$s^0_{42} = h_\theta(o_{42})$$
$$r^1_{42}, s^1_{42} = g_\theta(s^0_{42}, a_{42})$$
$$r^2_{42}, s^2_{42} = g_\theta(s^1_{42}, a_{42})$$

$$v^0_{42}, p^0_{42} = f_\theta(s^0_{42})$$
$$v^1_{42}, p^1_{42} = f_\theta(s^1_{42})$$
$$v^2_{42}, p^2_{42} = f_\theta(s^2_{42})$$

MuZero Predict several steps ahead

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}		
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}		

$$
\begin{aligned}
s^0_{42} &= h_\theta(o_{42}) \\
r^1_{42}, s^1_{42} &= g_\theta(s^0_{42}, a_{42}) \\
r^2_{42}, s^2_{42} &= g_\theta(s^1_{42}, a_{42}) \\
&\vdots
\end{aligned}
\qquad
\begin{aligned}
v^0_{42}, p^0_{42} &= f_\theta(s^0_{42}) \\
v^1_{42}, p^1_{42} &= f_\theta(s^1_{42}) \\
v^2_{42}, p^2_{42} &= f_\theta(s^2_{42})
\end{aligned}
$$

MuZero Predict several steps ahead

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$s^0_{42} = h_\theta(o_{42})$$
$$r^1_{42}, s^1_{42} = g_\theta(s^0_{42}, a_{42})$$
$$r^2_{42}, s^2_{42} = g_\theta(s^1_{42}, a_{42})$$
$$\vdots$$

$$v^0_{42}, p^0_{42} = f_\theta(s^0_{42})$$
$$v^1_{42}, p^1_{42} = f_\theta(s^1_{42})$$
$$v^2_{42}, p^2_{42} = f_\theta(s^2_{42})$$
$$\vdots$$

MuZero

Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

MuZero

Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

reward loss

$$l^r(u_{43}, r^1_{42})$$

MuZero

Reward loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^r(u_{43}, r^1_{42}) + l^r(u_{44}, r^2_{42})$$

MuZero

Reward loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^r(u_{43}, r^1_{42}) + l^r(u_{44}, r^2_{42}) + l^r(u_{45}, r^3_{42})$$

MuZero

Reward loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^r(u_{43}, r^1_{42}) + l^r(u_{44}, r^2_{42}) + l^r(u_{45}, r^3_{42}) + l^r(u_{46}, r^4_{42})$$

MuZero

Reward loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^r(u_{43}, r^1_{42}) + l^r(u_{44}, r^2_{42}) + l^r(u_{45}, r^3_{42}) + l^r(u_{46}, r^4_{42})$$

$$= \sum_{k=1}^{4} l^r(u_{42+k}, r^k_{42})$$

MuZero

Policy loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

MuZero

Policy loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^p(\pi_{42}, p^0_{42})$$

MuZero

Policy loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^p(\pi_{42}, p^0_{42}) + l^p(u_{43}, p^1_{42})$$

MuZero

Policy loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^p(\pi_{42}, p^0_{42}) + l^p(u_{43}, p^1_{42}) + l^p(u_{44}, p^2_{42})$$

MuZero

Policy loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^p(\pi_{42}, p^0_{42}) + l^p(u_{43}, p^1_{42}) + l^p(u_{44}, p^2_{42}) + l^p(\pi_{45}, p^3_{42})$$

MuZero

Policy loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^p(\pi_{42}, p^0_{42}) + l^p(u_{43}, p^1_{42}) + l^p(u_{44}, p^2_{42}) + l^p(\pi_{45}, p^3_{42}) + l^p(\pi_{46}, p^4_{42})$$

MuZero

Policy loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^p(\pi_{42}, p^0_{42}) + l^p(u_{43}, p^1_{42}) + l^p(u_{44}, p^2_{42}) + l^p(\pi_{45}, p^3_{42}) + l^p(\pi_{46}, p^4_{42})$$

$$= \sum_{k=0}^{4} l^p(\pi_{42+k}, p^k_{42})$$

MuZero

What is the target value?

Unroll steps

Time	Saved: State	Reward	Action	Improved Policy	Improved Value	Predicted/Calculated: Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

dont use as-is

Use n-step TD

MuZero What is the target value?

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

Use n-step TD

$$z_{42} = u_{43} + \gamma u_{44} + \gamma^2 u_{45} + \cdots + \gamma^{n-1} u_{42+n} + \gamma^n \nu_{42+n}$$

target value

MuZero What is the target value?

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy	
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}	z_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}	z_{43}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}	z_{44}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}	z_{45}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}	z_{46}

Use n-step TD

n-step lookahead

$$z_{42} = u_{43} + \gamma u_{44} + \gamma^2 u_{45} + \cdots + \gamma^{n-1} u_{42+n} + \gamma^n \nu_{42+n}$$

$$z_{43} = u_{44} + \gamma u_{45} + \gamma^2 u_{46} + \cdots + \gamma^{n-1} u_{43+n} + \gamma^n \nu_{43+n}$$

n-step lookahead

MuZero

What is the target value?

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

Use n-step TD

$$z_{42} = u_{43} + \gamma u_{44} + \gamma^2 u_{45} + \cdots + \gamma^{n-1} u_{42+n} + \gamma^n \nu_{42+n}$$
$$z_{43} = u_{44} + \gamma u_{45} + \gamma^2 u_{46} + \cdots + \gamma^{n-1} u_{43+n} + \gamma^n \nu_{43+n}$$
$$z_t = u_{t+1} + \gamma u_{t+2} + \gamma^2 u_{t+3} + \cdots + \gamma^{n-1} u_{t+n} + \gamma^n \nu_{t+n}$$

MuZero

Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

MuZero

Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^v(z_{42}, v^0_{42})$$

MuZero

Value loss

Unroll steps

Time	Saved: State	Saved: Reward	Saved: Action	Saved: Improved Policy	Saved: Improved Value	Predicted/Calculated: Reward	Predicted/Calculated: Value	Predicted/Calculated: Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

z_{42} z_{43} ... $z_?$

$$l^v(z_{42}, v^0_{42}) + l^v(z_{43}, v^1_{42})$$

MuZero

Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^v(z_{42}, v^0_{42}) + l^v(z_{43}, v^1_{42}) + l^v(z_{44}, v^2_{42})$$

MuZero

Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^v(z_{42}, v^0_{42}) + l^v(z_{43}, v^1_{42}) + l^v(z_{44}, v^2_{42}) + l^v(z_{45}, v^3_{42})$$

MuZero

Value loss

Time	Saved: State	Reward	Action	Improved Policy	Improved Value	Predicted/Calculated: Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

Unroll steps

z_{46}

v_{47} v_{48} v_{49}

$$l^v(z_{42}, v^0_{42}) + l^v(z_{43}, v^1_{42}) + l^v(z_{44}, v^2_{42}) + l^v(z_{45}, v^3_{42}) + l^v(z_{46}, v^4_{42})$$

MuZero

Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	o_{42}	u_{42}	a_{42}	π_{42}	ν_{42}		v^0_{42}	p^0_{42}
43		u_{43}	a_{43}	π_{43}	ν_{43}	r^1_{42}	v^1_{42}	p^1_{42}
44		u_{44}	a_{44}	π_{44}	ν_{44}	r^2_{42}	v^2_{42}	p^2_{42}
45		u_{45}	a_{45}	π_{45}	ν_{45}	r^3_{42}	v^3_{42}	p^3_{42}
46		u_{46}	a_{46}	π_{46}	ν_{46}	r^4_{42}	v^4_{42}	p^4_{42}

$$l^v(z_{42}, v^0_{42}) + l^v(z_{43}, v^1_{42}) + l^v(z_{44}, v^2_{42}) + l^v(z_{45}, v^3_{42}) + l^v(z_{46}, v^4_{42})$$

$$= \sum_{k=0}^{4} l^v(z_{42+k}, v^k_{42})$$

MuZero

Total loss

$$l_t(\theta) = \sum_{k=0}^{K} l^r(u_{t+k}, r_t^k) + l^v(z_{t+k}, v_t^k) + l^p(\pi_{t+k}, p_t^k) + c \parallel \theta \parallel^2$$

encourages small network weights

- Sum of:
  - Reward loss
  - Policy loss
  - Value loss
  - Regularization term to prevent overfitting

MuZero

Boardgames vs. Atari Games

$z_0 = z_1 = z_{42}$ = final game outcome

	TD	Unroll steps	Gamma	MCTS Simulations per move
Boardgames	TD(∞)	5	1 (no discount)	800
Atari Games	TD(10)	5	0.997	50

how far ahead do we check loss on our dynamics

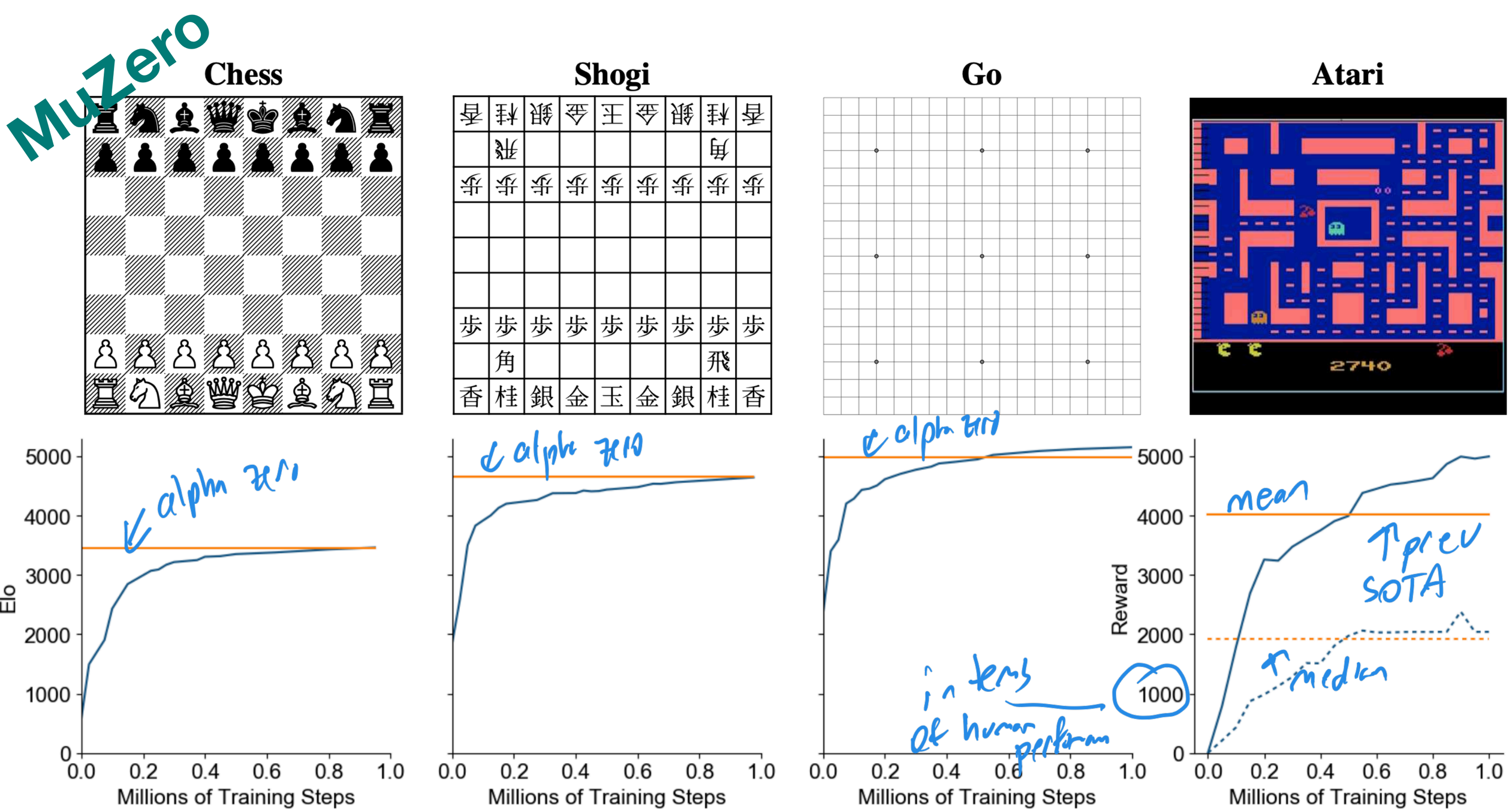


Figure 2: **Evaluation of *MuZero* throughout training in chess, shogi, Go and Atari.** The x-axis shows millions of training steps. For chess, shogi and Go, the y-axis shows Elo rating, established by playing games against *AlphaZero* using 800 simulations per move for both players. *MuZero*'s Elo is indicated by the blue line, *AlphaZero*'s Elo by the horizontal orange line. For Atari, mean (full line) and median (dashed line) human normalized scores across all 57 games are shown on the y-axis. The scores for R2D2 [21], (the previous state of the art in this domain, based on model-free RL) are indicated by the horizontal orange lines. Performance in Atari was evaluated using 50 simulations every fourth time-step, and then repeating the chosen action four times, as in prior work [25].

Summary

RL consists of

- an agent
- interacting with an environment
- receiving rewards

(assum MDP)

Goal of agent:

- Learn policy to maximize expected long-term (discounted) reward

Summary