

# 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-python-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

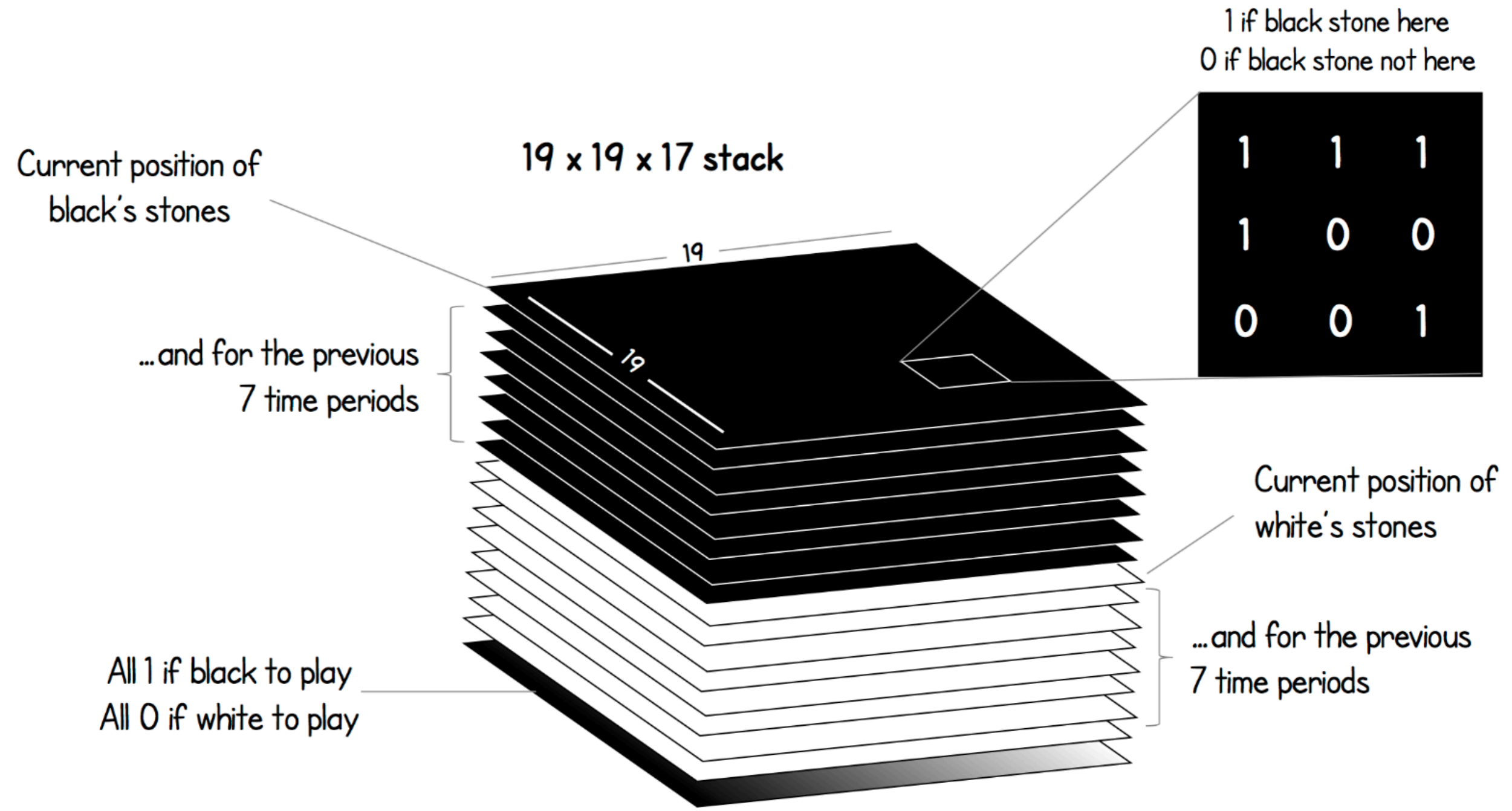
- AlphaGo Zero, 2017
  - No prior knowledge of Go (no hand-crafted features, no training on human games). All self-play
  - Single Residual Neural Network learns both value and policy functions
  - Keeps best-neural-network-so-far (new champion if beats  $>55\%$  of games against old champion)

# 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!

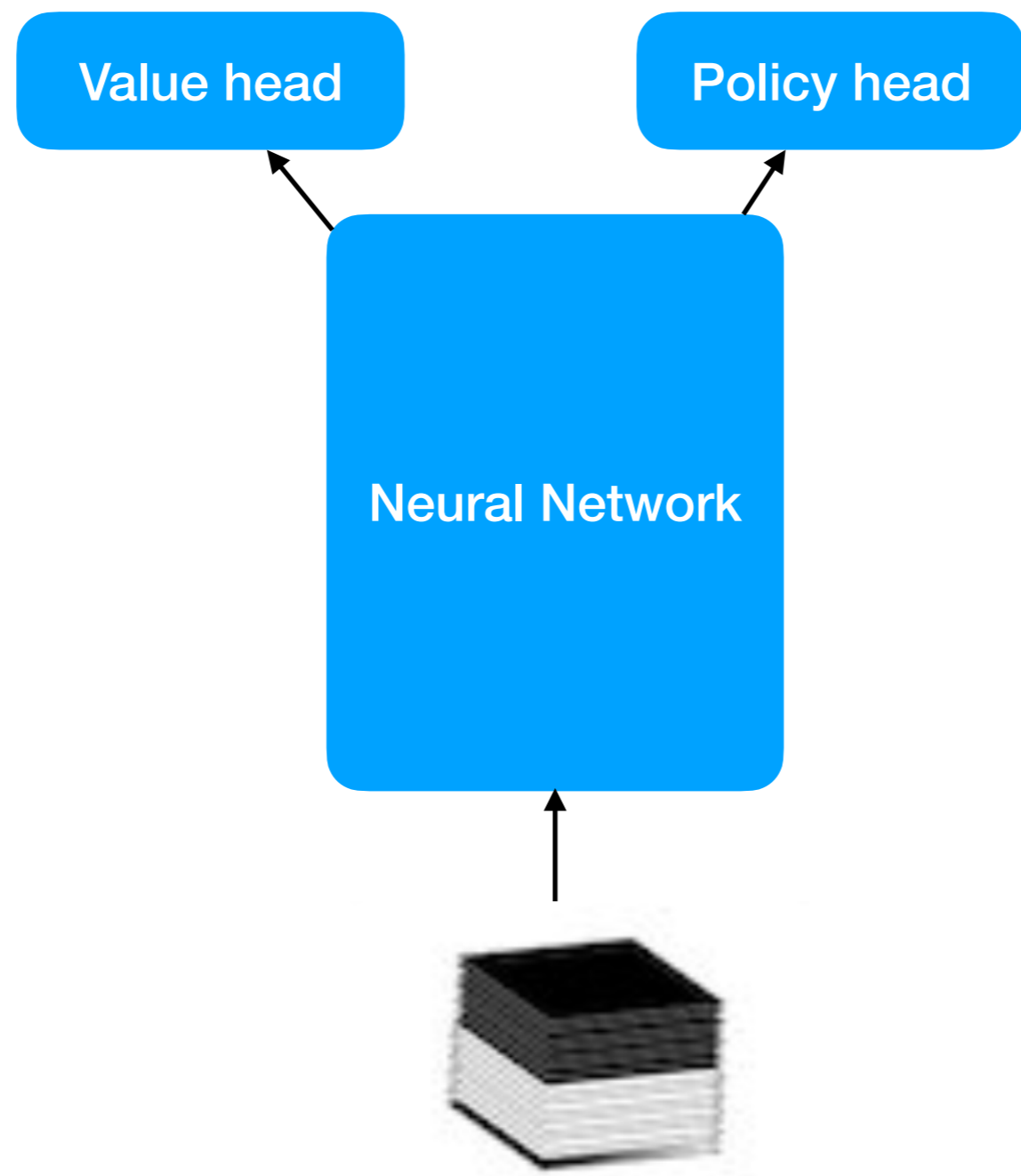
AlphaGo  
Zero

# AlphaGo Zero: What is a game state for Go?



AlphaGo  
Zero

# The Neural Network

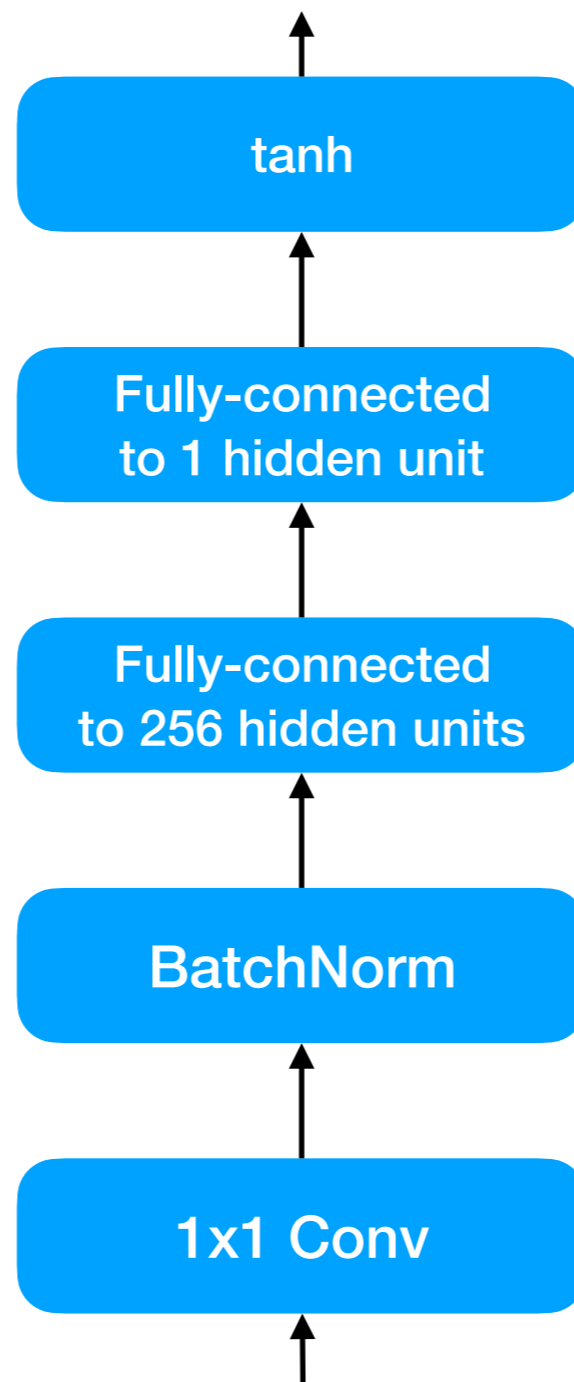


Multi-task learning!

Input (Game state)

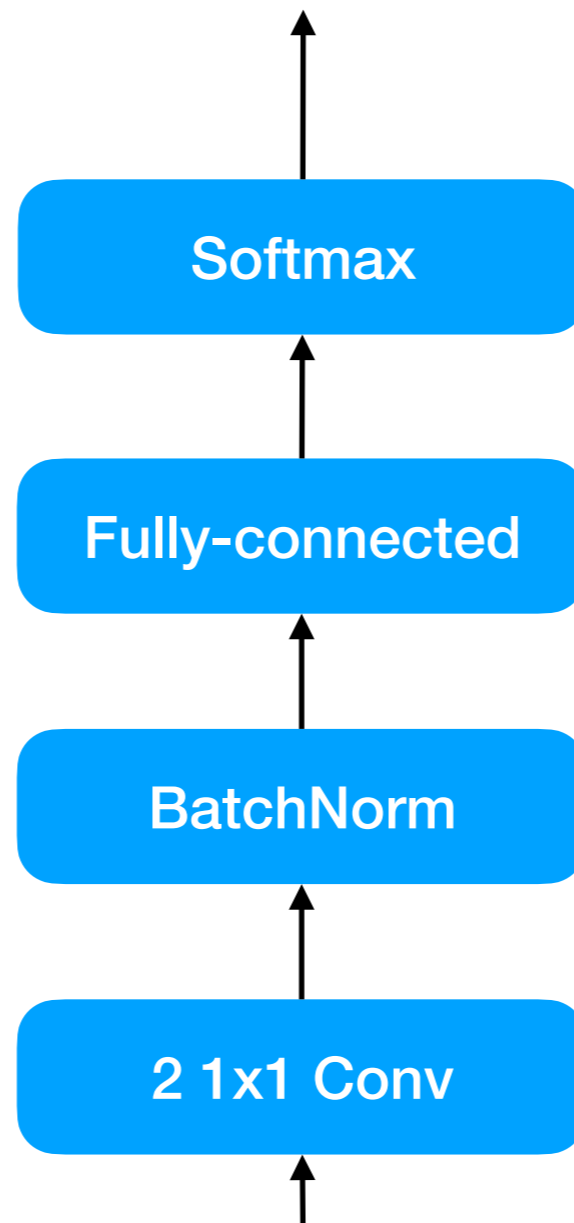
# Value (V) Head

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



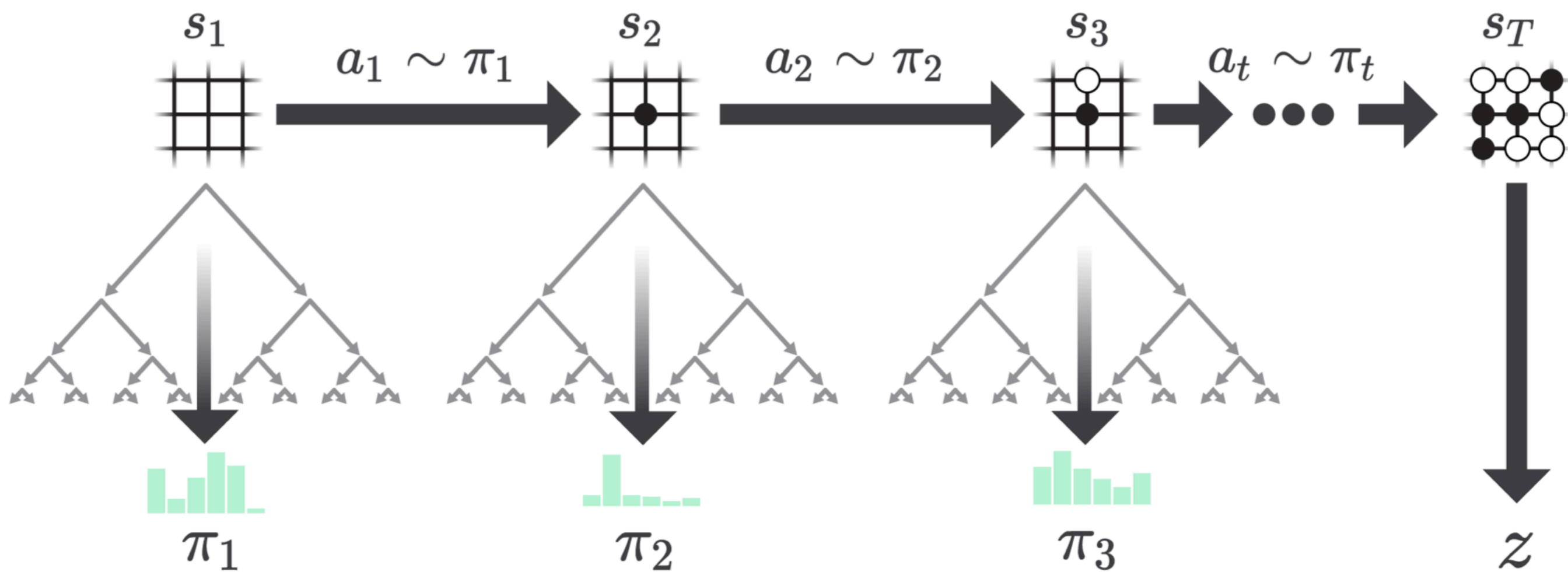
# Policy (P) Head

19x19+1 probabilities

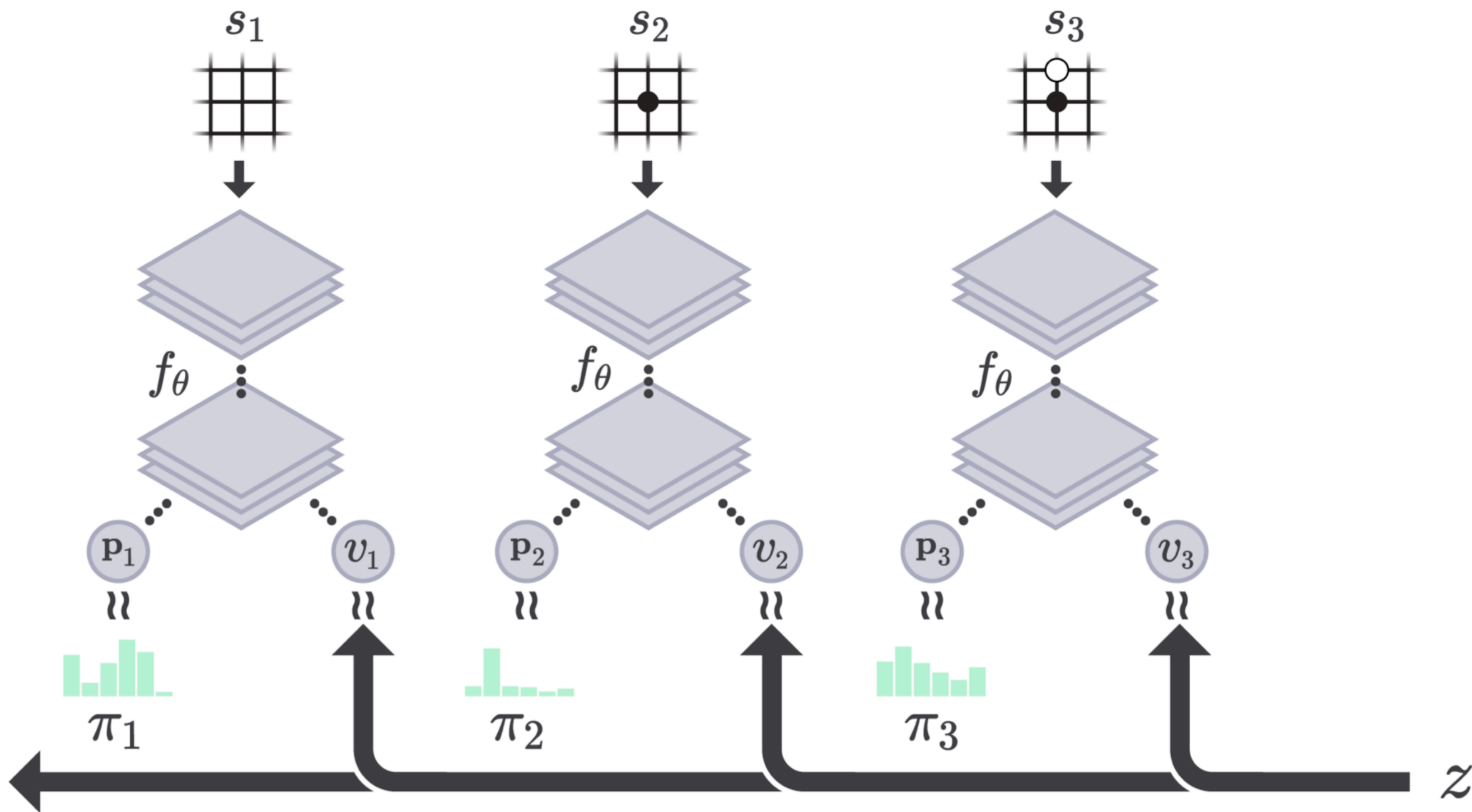


19x19 places to play stone  
1 way to pass

# Self-play



# Neural Net Training



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

# Neural Net Training

- Update  $\Theta$  so that:
  - Given an input state,  $s_i$ ,  $f_{\Theta}$  produces output that is closer to:
    - For value portion:  $z$ , the actual result of the game
    - For action portion:  $\pi_i$ , the MCTS-improved policy based on the neural net's  $p_i$

# 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:  $W(s, a)/N(s, a)$**

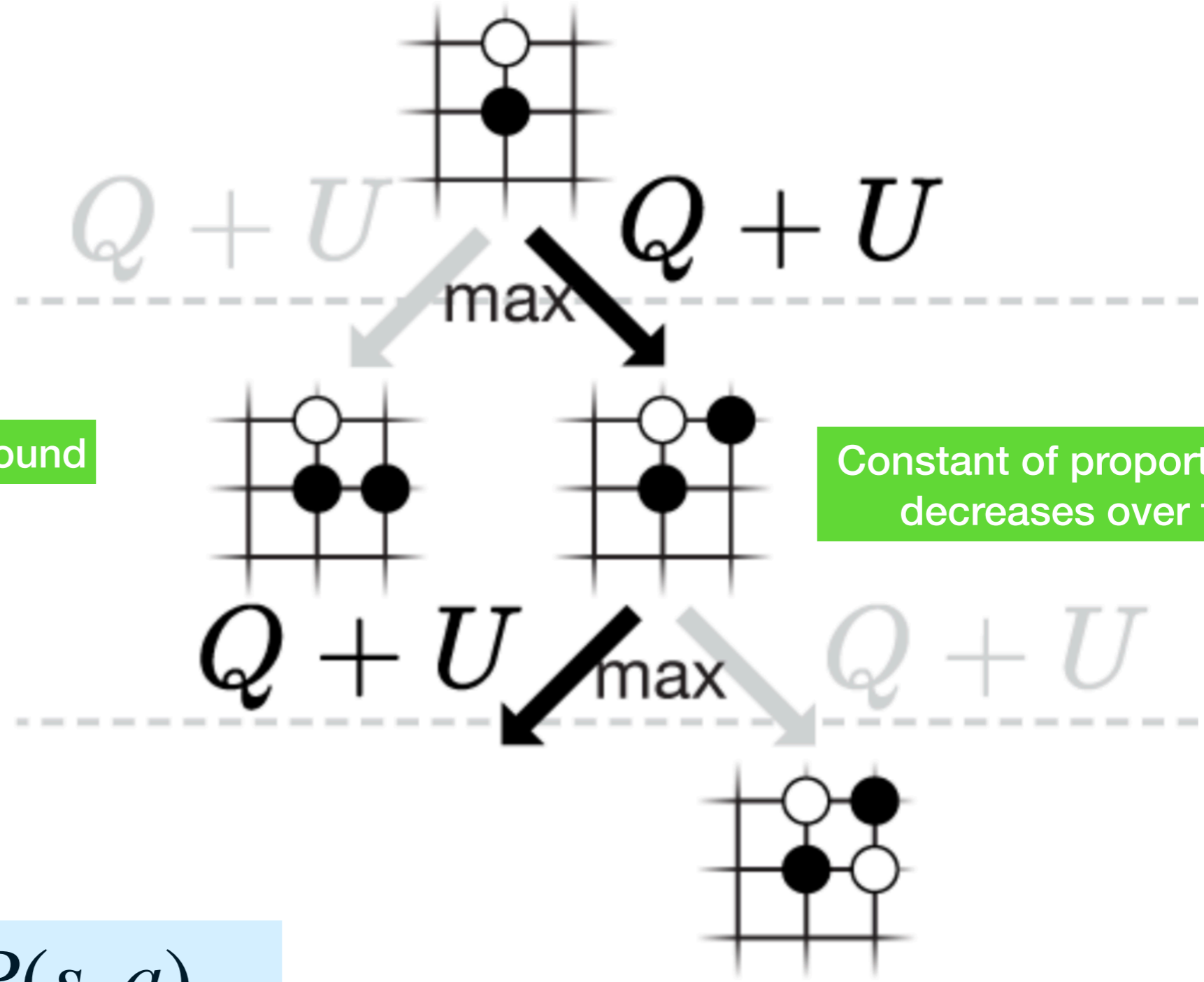
**$P(s, a)$ : prior probability of selecting that edge**

# 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

# MCTS: Select

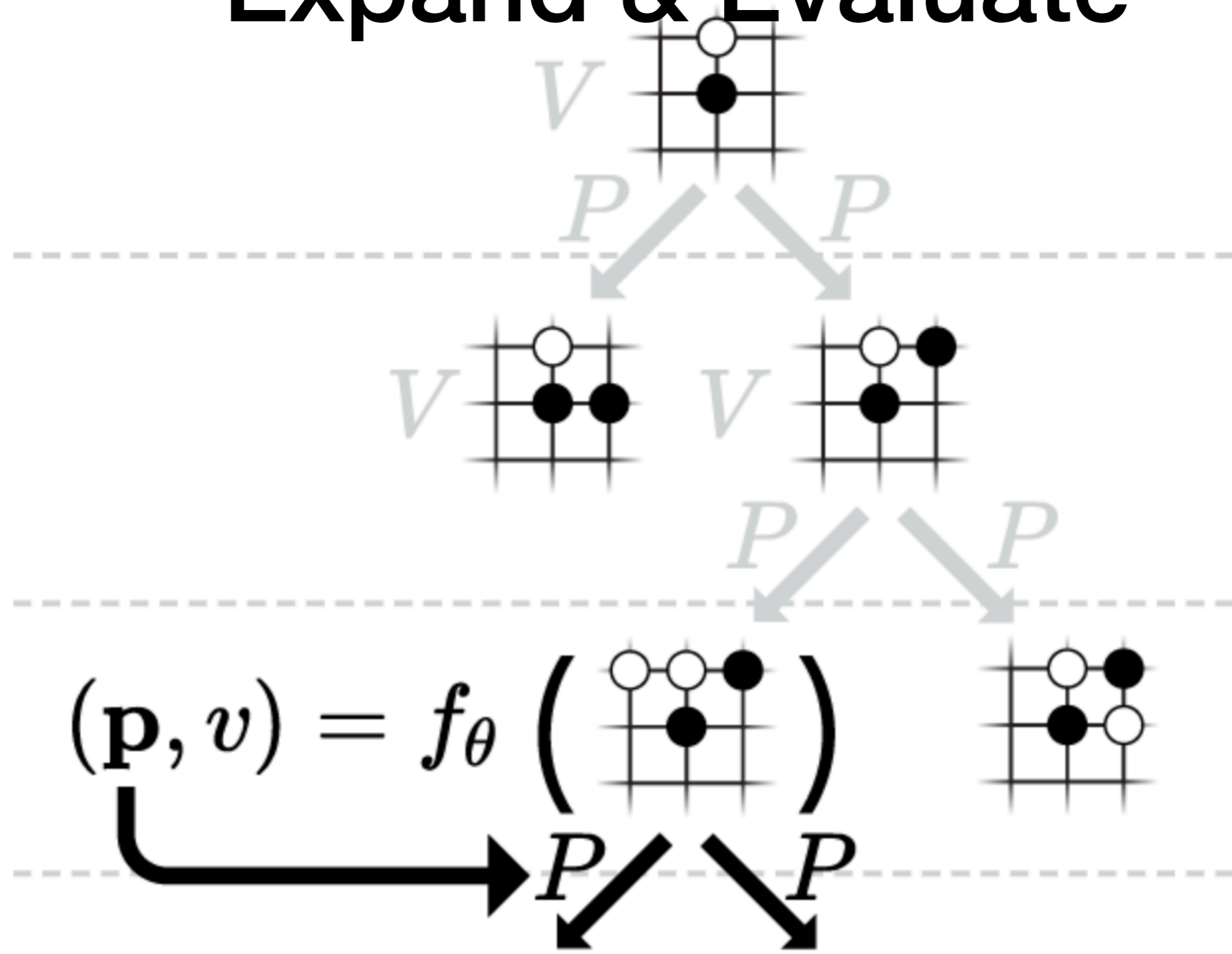


U is upper-confidence bound

Constant of proportionality decreases over time

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

# MCTS: Expand & Evaluate

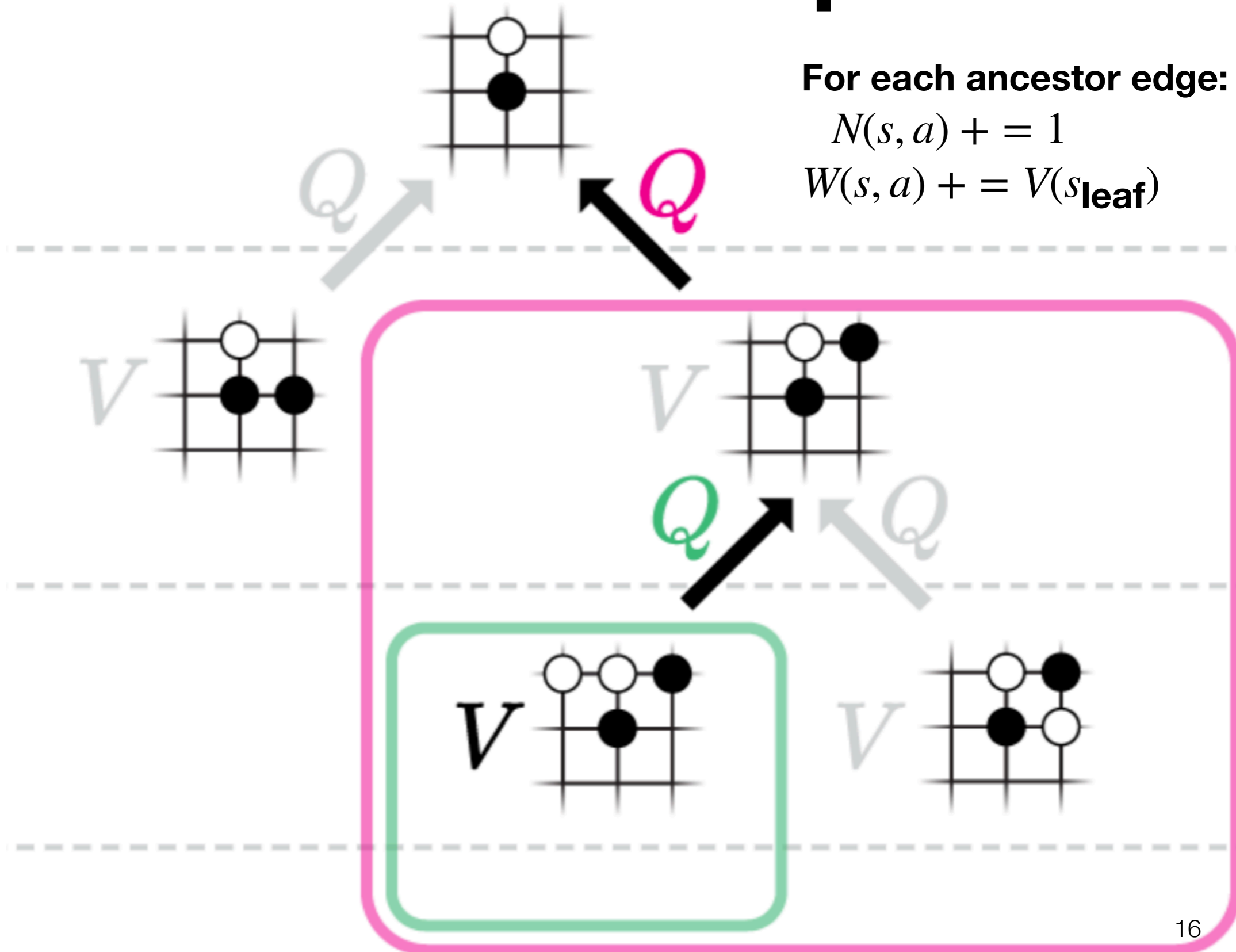


# MCTS: Backup

For each ancestor edge:

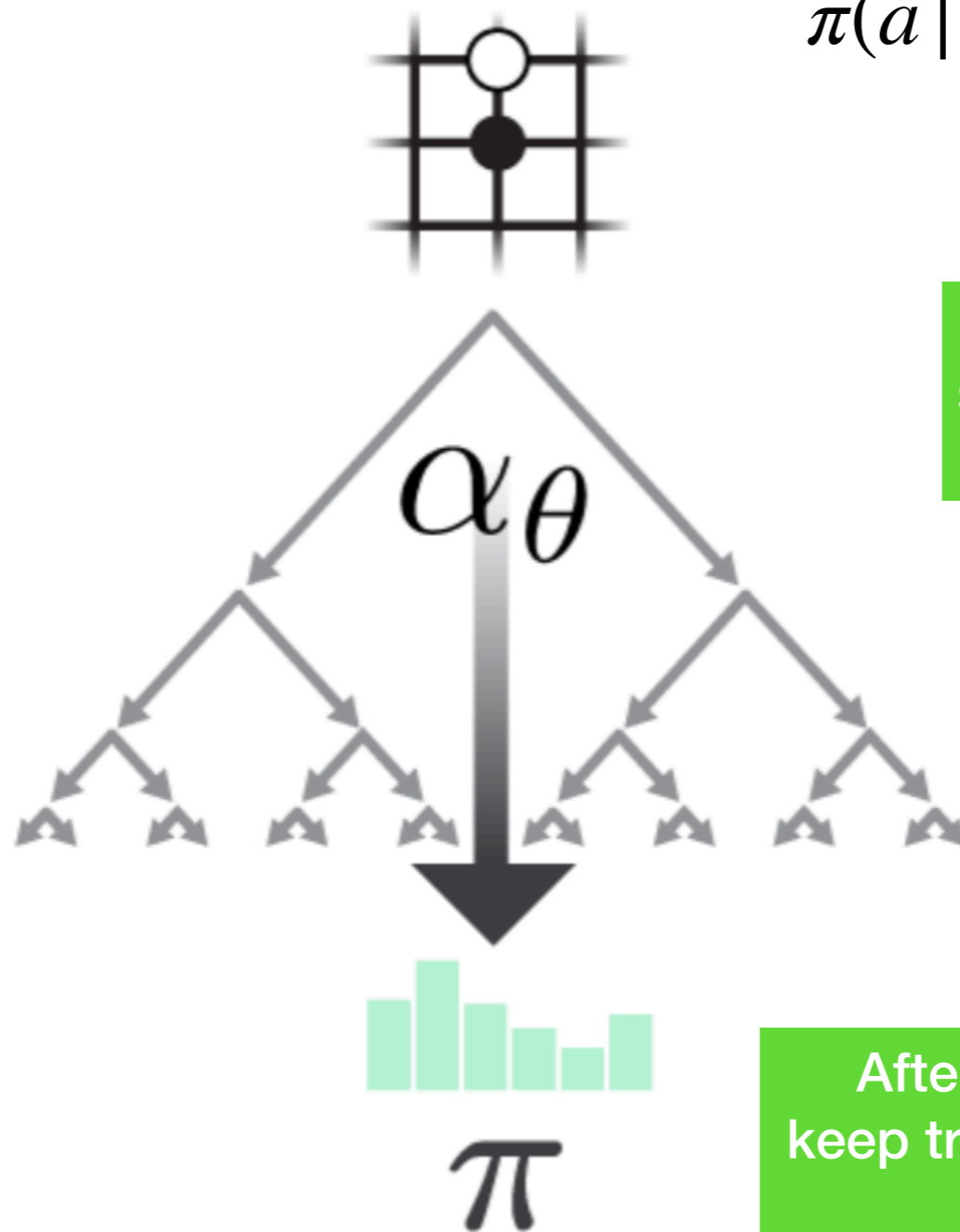
$$N(s, a) += 1$$

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



# MCTS: Play

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



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

After an action is chosen from  $\pi$ , keep tree starting from resulting state, clear rest of tree

# What Go-specific knowledge is used?

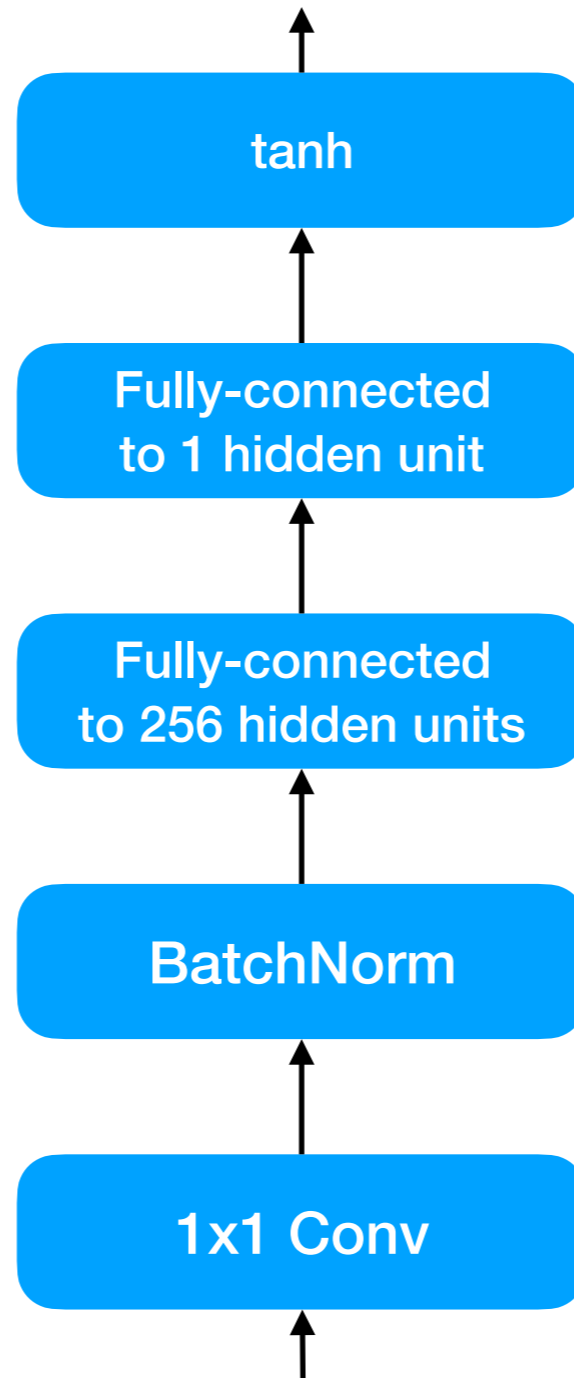
- Network knows input format (19x19 board, etc.)
- Network knows how many possible actions ( $19 \times 19 + 1$ )
- MCTS knows which actions lead to which states
- MCTS knows whether a state is terminal
- MCTS knows how to score a terminal state
- MCTS does dihedral reflection or rotation (takes advantage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

# 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

# Value (V) Head

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



# 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

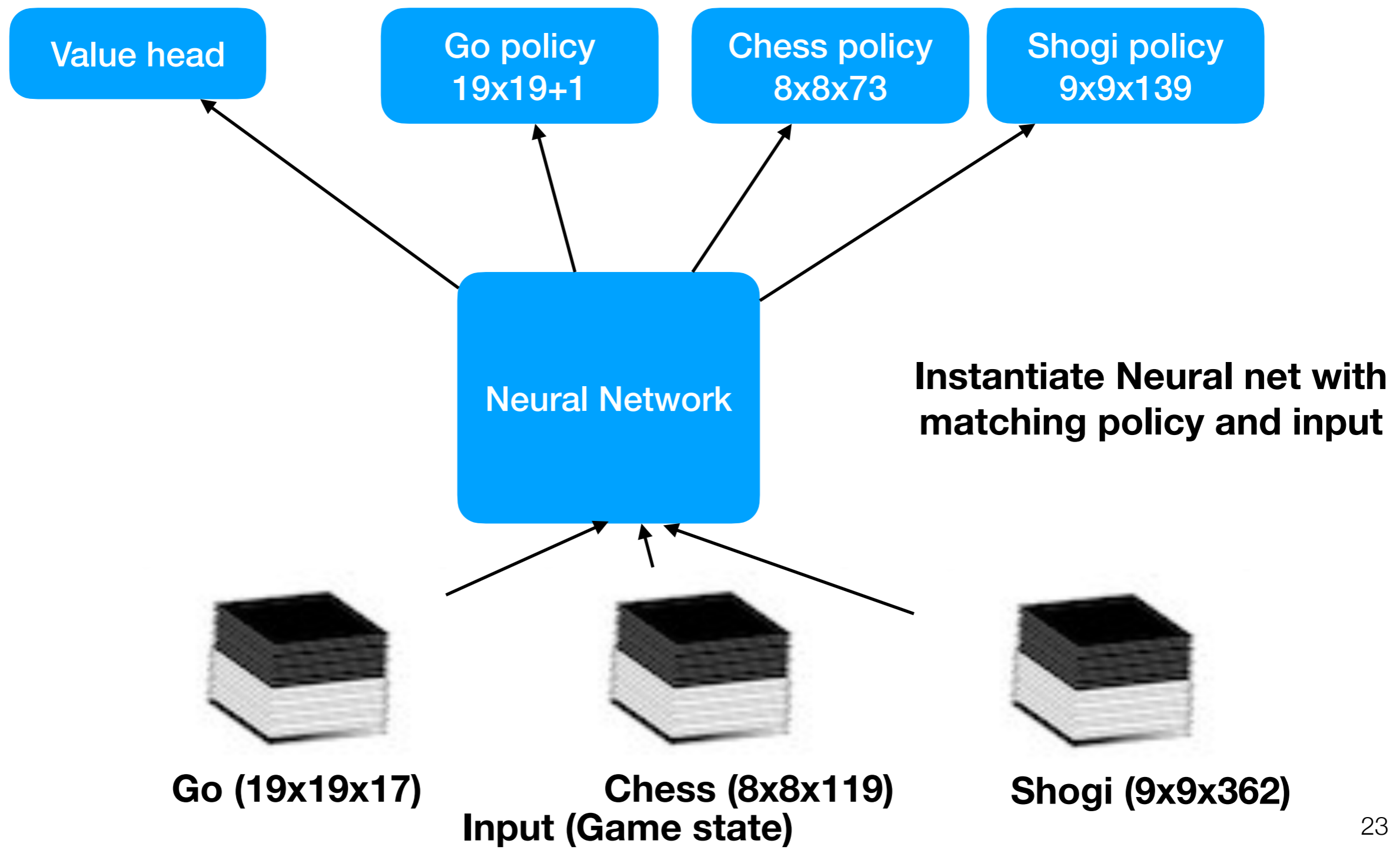
# What is a game state for Chess?

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

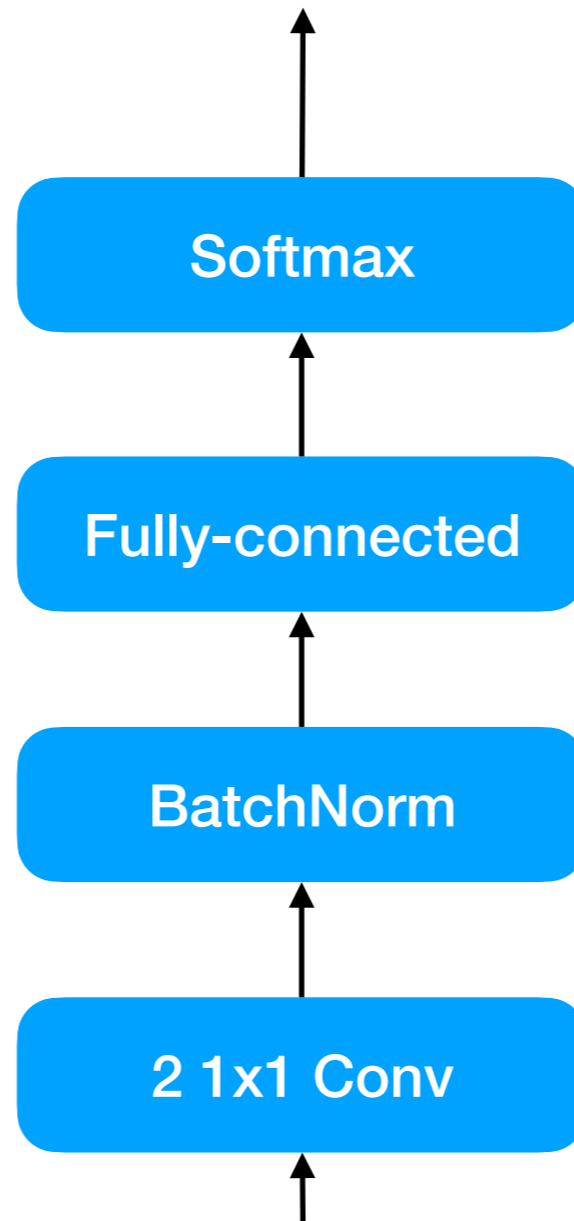
+ Extra planes for: repetition, color, total move count, and a few others

# The Neural Network



# Policy (P) Head (for Chess)

8x8x73 probabilities



8x8 for picking up a piece  
56 for 8 directions (N, NE, ...)  
x 7 distance  
8 for 8 possible knight moves  
9 = 3x3: 3 underpromotions (K/R/B)  
with 3 ways to get to 8th place  
(straight, capture diagonal L/R)

# Results

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

# Takeaways

- Simple representation of input and output
- MCTS improves  $P$  to  $\pi$
- Train NN:  $x = \text{game state}$ ,  $y = (\pi, \text{game\_result})$ ,  $\hat{y} = (P, V)$

# 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

# Planning with a Learned Model

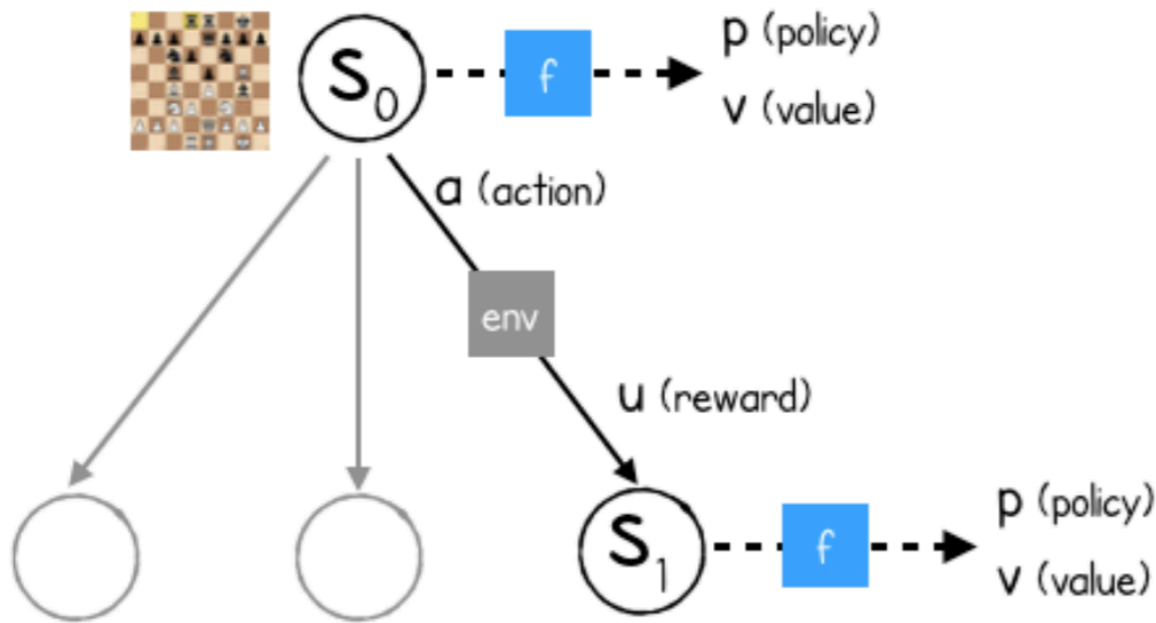
- Network knows input format
- Network knows how many possible actions
- ~~MCTS knows which actions lead to which states~~
- ~~MCTS knows whether a state is terminal~~
- ~~MCTS knows how to score a terminal state~~
- Typical number of legal moves scales exploration noise

# 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 ( $s_k$ )
  - Train jointly three functions (NN):
    - representation  $h_\theta : o_k \rightarrow s_k$
    - dynamics  $g_\theta : s_k, a_k \rightarrow r_{k+1}, s_{k+1}$
    - prediction  $f_\theta : s_k \rightarrow p_k, v_k$

MuZero

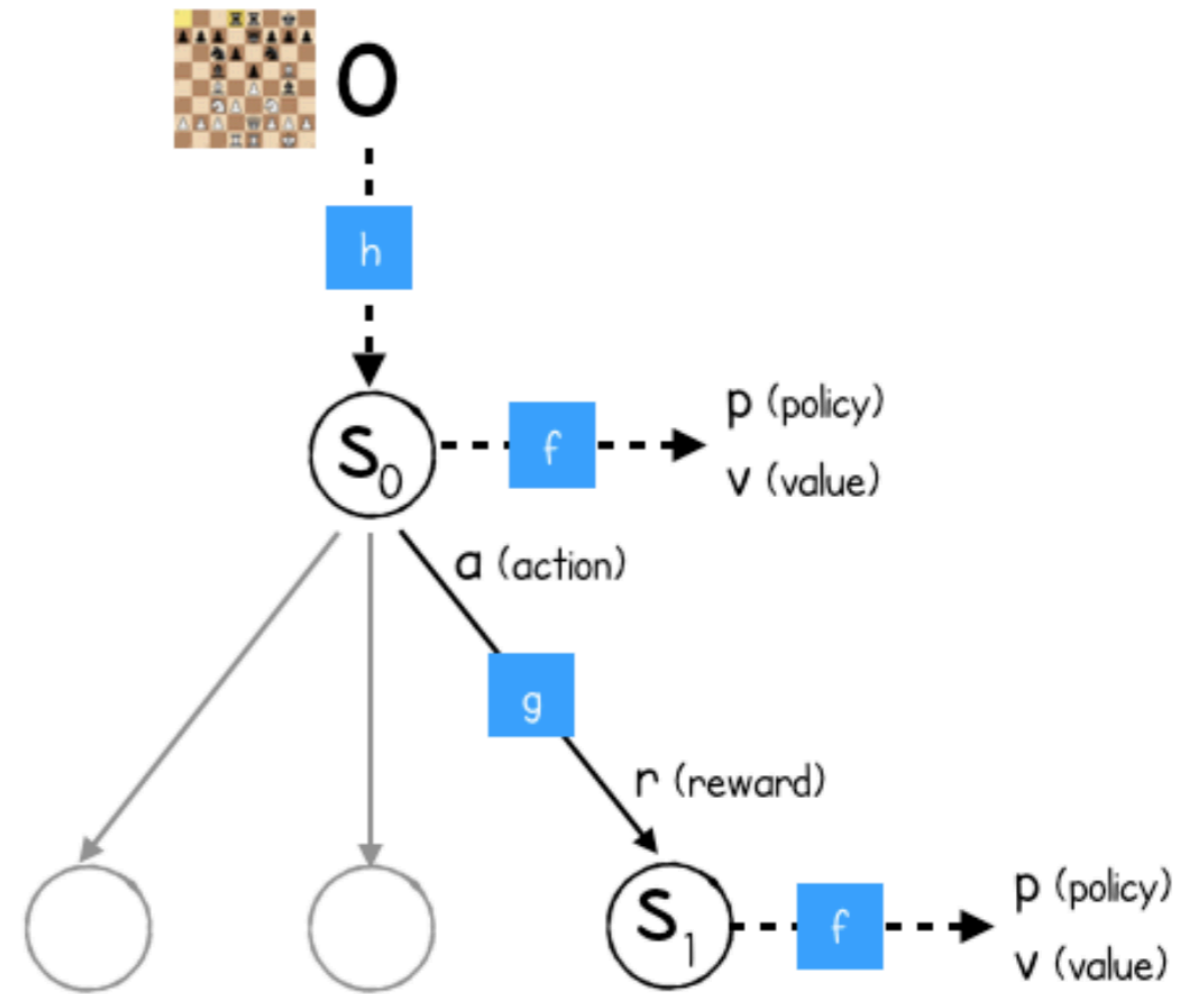
# AlphaZero



AlphaZero has 1 network

	from	to
prediction	$f: s$	$p, v$

# MuZero

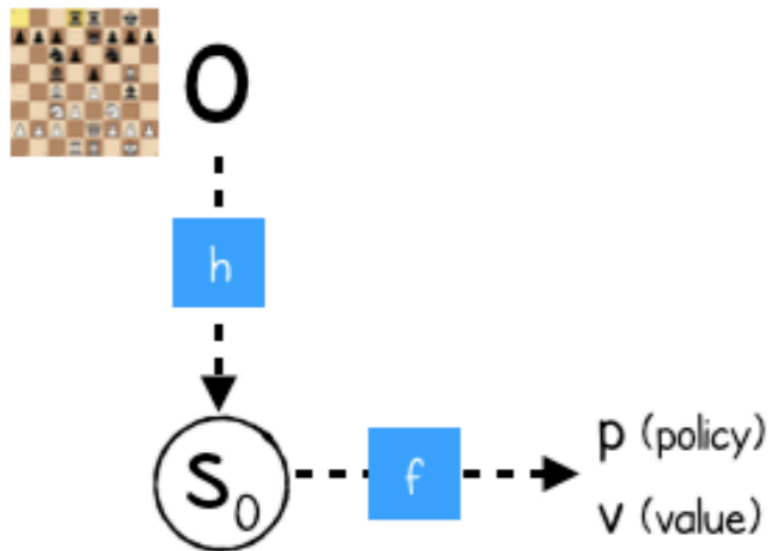


MuZero has 3 networks

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

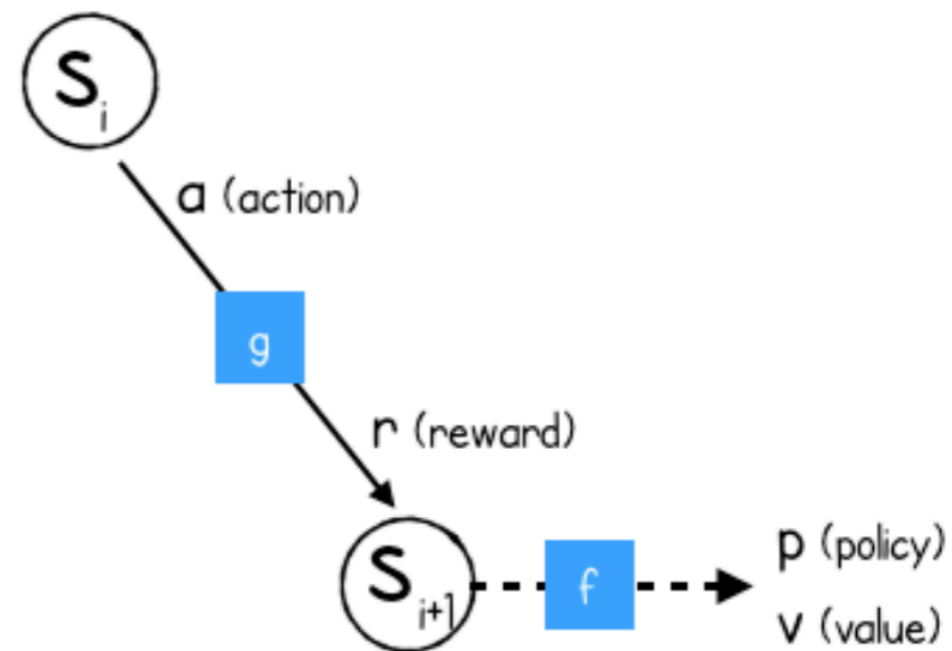
# MuZero

initial\_inference



<u>input</u>	<u>output</u>
image o	value v
	reward r (=0)
	policy_logits p
	hidden_state s <sub>0</sub>

recurrent\_inference



<u>input</u>	<u>output</u>
hidden_state s <sub>i</sub>	value v
action a	reward r
	policy_logits p
	hidden_state s <sub>i+1</sub>

# MCTS:

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

# 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

## Game 5, observation 42

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$			
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$			
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$			
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

Unroll steps

# Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$			
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$			
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$			
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

# MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$			
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$			
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$			
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

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

# MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$			
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$			
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

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

$$v_{42}^0, p_{42}^0 = f_{\theta}(s_{42}^0)$$

# MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$		
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$			
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

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

$$r_{42}^1, s_{42}^1 = g_{\theta}(s_{42}^0, a_{42})$$

$$v_{42}^0, p_{42}^0 = f_{\theta}(s_{42}^0)$$

# MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$			
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

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

$$r_{42}^1, s_{42}^1 = g_{\theta}(s_{42}^0, a_{42})$$

$$v_{42}^0, p_{42}^0 = f_{\theta}(s_{42}^0)$$

$$v_{42}^1, p_{42}^1 = f_{\theta}(s_{42}^1)$$

# Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$		
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

$$\begin{aligned}
 s_{42}^0 &= h_{\theta}(o_{42}) & v_{42}^0, p_{42}^0 &= f_{\theta}(s_{42}^0) \\
 r_{42}^1, s_{42}^1 &= g_{\theta}(s_{42}^0, a_{42}) & v_{42}^1, p_{42}^1 &= f_{\theta}(s_{42}^1) \\
 r_{42}^2, s_{42}^2 &= g_{\theta}(s_{42}^1, a_{42}) & &
 \end{aligned}$$

# Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$			
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$			

$$\begin{aligned}
 s_{42}^0 &= h_{\theta}(o_{42}) & v_{42}^0, p_{42}^0 &= f_{\theta}(s_{42}^0) \\
 r_{42}^1, s_{42}^1 &= g_{\theta}(s_{42}^0, a_{42}) & v_{42}^1, p_{42}^1 &= f_{\theta}(s_{42}^1) \\
 r_{42}^2, s_{42}^2 &= g_{\theta}(s_{42}^1, a_{42}) & v_{42}^2, p_{42}^2 &= f_{\theta}(s_{42}^2)
 \end{aligned}$$

# MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$		
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$		

$$\begin{aligned}
 s_{42}^0 &= h_{\theta}(o_{42}) & v_{42}^0, p_{42}^0 &= f_{\theta}(s_{42}^0) \\
 r_{42}^1, s_{42}^1 &= g_{\theta}(s_{42}^0, a_{42}) & v_{42}^1, p_{42}^1 &= f_{\theta}(s_{42}^1) \\
 r_{42}^2, s_{42}^2 &= g_{\theta}(s_{42}^1, a_{42}) & v_{42}^2, p_{42}^2 &= f_{\theta}(s_{42}^2) \\
 &\vdots & &
 \end{aligned}$$

# MuZero Predict several steps ahead

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$o_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

$$\begin{aligned}
 s_{42}^0 &= h_{\theta}(o_{42}) \\
 r_{42}^1, s_{42}^1 &= g_{\theta}(s_{42}^0, a_{42}) \\
 r_{42}^2, s_{42}^2 &= g_{\theta}(s_{42}^1, a_{42}) \\
 &\vdots
 \end{aligned}$$

$$\begin{aligned}
 v_{42}^0, p_{42}^0 &= f_{\theta}(s_{42}^0) \\
 v_{42}^1, p_{42}^1 &= f_{\theta}(s_{42}^1) \\
 v_{42}^2, p_{42}^2 &= f_{\theta}(s_{42}^2) \\
 &\vdots
 \end{aligned}$$

# Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

# Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Reward loss

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Reward loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

$$\begin{aligned}
 & l^r(u_{43}, r_{42}^1) + l^r(u_{44}, r_{42}^2) + l^r(u_{45}, r_{42}^3) + l^r(u_{46}, r_{42}^4) \\
 &= \sum_{k=1}^4 l^r(u_{42+k}, r_{42}^k)
 \end{aligned}$$

# Policy loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

# Policy loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Policy loss

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Policy loss

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Policy loss

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Policy loss

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Policy loss

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

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

# What is the target value?

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

Unroll steps

Use n-step TD

# MuZero What is the target value?

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

Use n-step TD

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

# What is the target value?

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

Unroll steps

## Use n-step TD

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

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

# What is the target value?

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

Unroll steps

## Use n-step TD

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

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

$$z_t = u_{t+1} + \gamma u_{t+2} + \gamma^2 u_{t+3} + \dots + \gamma^{n-1} u_{t+n} + \gamma^n v_{t+n}$$

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

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

# Value loss

Unroll steps

Time	Saved					Predicted/Calculated		
	State	Reward	Action	Improved Policy	Improved Value	Reward	Value	Policy
42	$O_{42}$	$u_{42}$	$a_{42}$	$\pi_{42}$	$v_{42}$		$v_{42}^0$	$p_{42}^0$
43		$u_{43}$	$a_{43}$	$\pi_{43}$	$v_{43}$	$r_{42}^1$	$v_{42}^1$	$p_{42}^1$
44		$u_{44}$	$a_{44}$	$\pi_{44}$	$v_{44}$	$r_{42}^2$	$v_{42}^2$	$p_{42}^2$
45		$u_{45}$	$a_{45}$	$\pi_{45}$	$v_{45}$	$r_{42}^3$	$v_{42}^3$	$p_{42}^3$
46		$u_{46}$	$a_{46}$	$\pi_{46}$	$v_{46}$	$r_{42}^4$	$v_{42}^4$	$p_{42}^4$

$$\begin{aligned}
 & l^v(z_{42}, v_{42}^0) + l^v(z_{43}, v_{42}^1) + l^v(z_{44}, v_{42}^2) + l^v(z_{45}, v_{42}^3) + l^v(z_{46}, v_{42}^4) \\
 &= \sum_{k=0}^4 l^v(z_{42+k}, v_{42}^k)
 \end{aligned}$$

# Total loss

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

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

# Boardgames vs. Atari Games

	TD	Unroll steps	Gamma	MCTS Simulations per move
Boardgames	TD( $\infty$ )	5	1	800
Atari Games	TD(10)	5	0.997	50

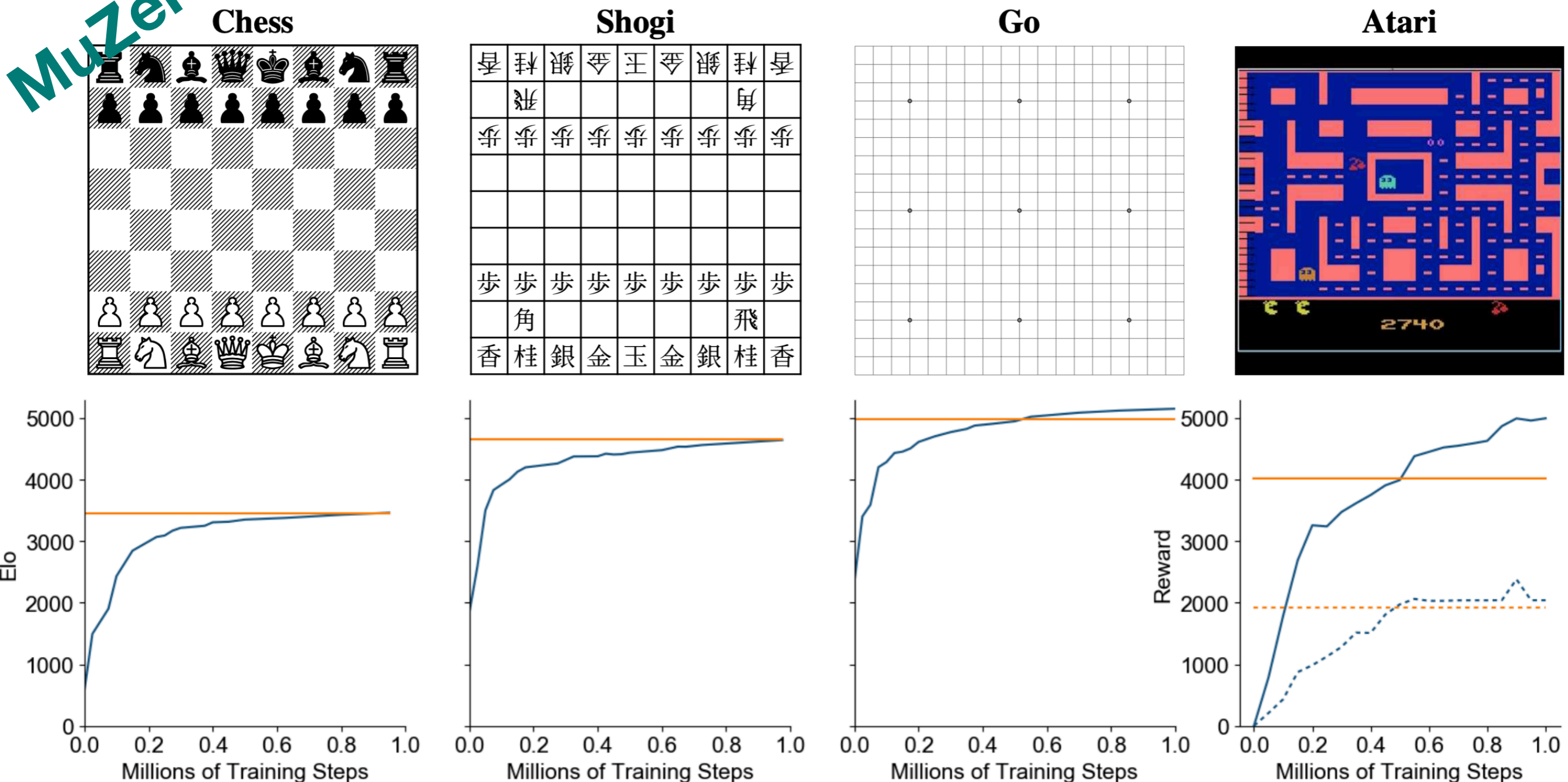


Figure 2: **Evaluation of *MuZero* throughout training in chess, shogi, Go and Atari.** The x-axis shows millions of training steps. For chess, shogi and Go, the y-axis shows Elo rating, established by playing games against *AlphaZero* using 800 simulations per move for both players. *MuZero*'s Elo is 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**

**Goal of agent:**

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

# Summary