DeepMind

Muesli: Combining Improvements in Policy Optimization

Matteo Hessel, Ivo Danihelka, Fabio Viola, Arthur Guez, Simon Schmitt, Laurent Sifre, Theophane Weber, David Silver, Hado van Hasselt

Benefits

- Make policy gradients robust to off-policy data and reward scales.
- Obtain MuZero's state-of-the-art score on Atari, even without MCTS.

Outline

1. Making policy gradients robust
2. The combined agent: "Muesli"
3. Results on Atari and 9x9 Go

Policy Gradients

With a function approximation,

the following properties are important:

- Able to learn a stochastic policy.
- Able to learn from an n-step return.
- Directly optimizing the acting. (Not depending on accurate models.)

The objective

Maximize the value from a start state.

$$v_\pi(s_0)$$

The sum of discounted rewards when following the policy π

Policy Gradient Theorem

Distribution of states

Action-value

$$\frac{\partial v_\pi(s_0)}{\partial \theta} = \sum_s d_\pi(s) \sum_a \frac{\partial \pi(a|s)}{\partial \theta} q_\pi(s, a)$$

Policy parameters.

[Sutton et al. 2000]

Policy Gradient Theorem

Distribution of states

Action-value

$$\frac{\partial v_\pi(s_0)}{\partial \theta} = \sum_s d_{\pi}(s) \sum_a \frac{\partial \pi(a|s)}{\partial \theta} q_{\pi}(s, a)$$

Policy parameters.

Violated by
starting the episode with π_{old}

Violated by
bootstrapping from $\hat{v}_{\pi_{\text{old}}}(s)$

Is it a problem?

The problem from policy mismatch

- The possible degradation of the policy value is related to a distance between π and π_{old}.

Policy loss

A regularizer.

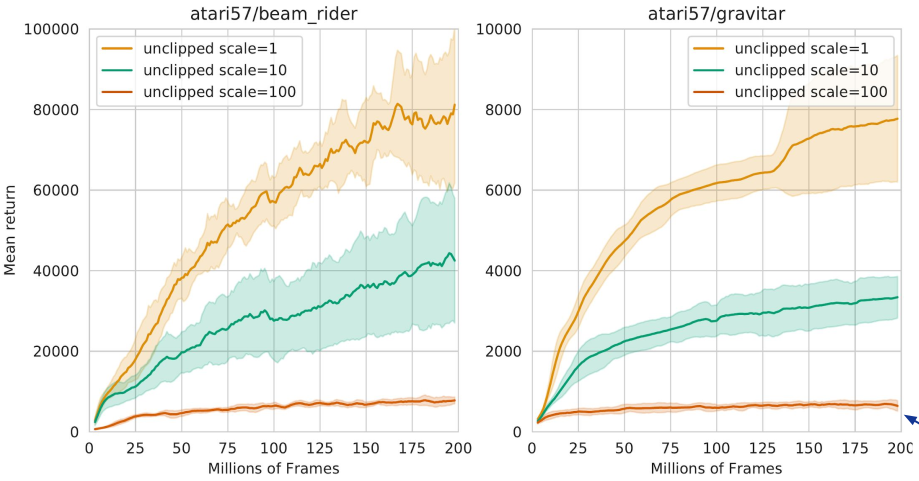
$$L(s, \pi) = L_{\mathrm{PG}}(s, \pi) + \mathrm{KL}(\pi_{\mathrm{CMPO}}, \pi)$$

Usual policy gradients.

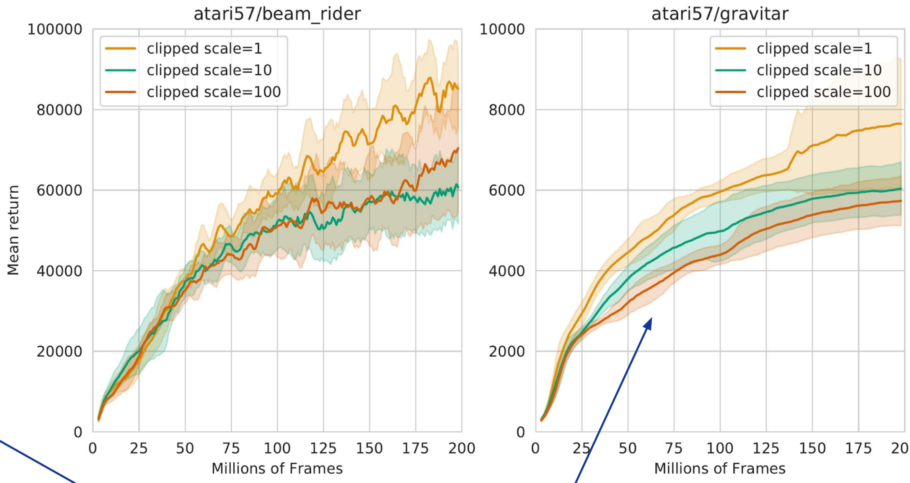
An improved policy, not too far from π_{old}.
The improved policy is constructed by
MPO with **clipped advantages**.

Clipped advantages are robust

With unclipped MPO advantages



With clipped MPO advantages



Scaling the advantages by **100**

Related work

- A natural policy gradient … clipped advantages = clipped update to policy logits.
- Conservative policy iteration
- Trust Region Policy Optimization (TRPO)
- Monte-Carlo Tree Search as regularized policy optimization
- Mirror Descent Policy Optimization
- Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning

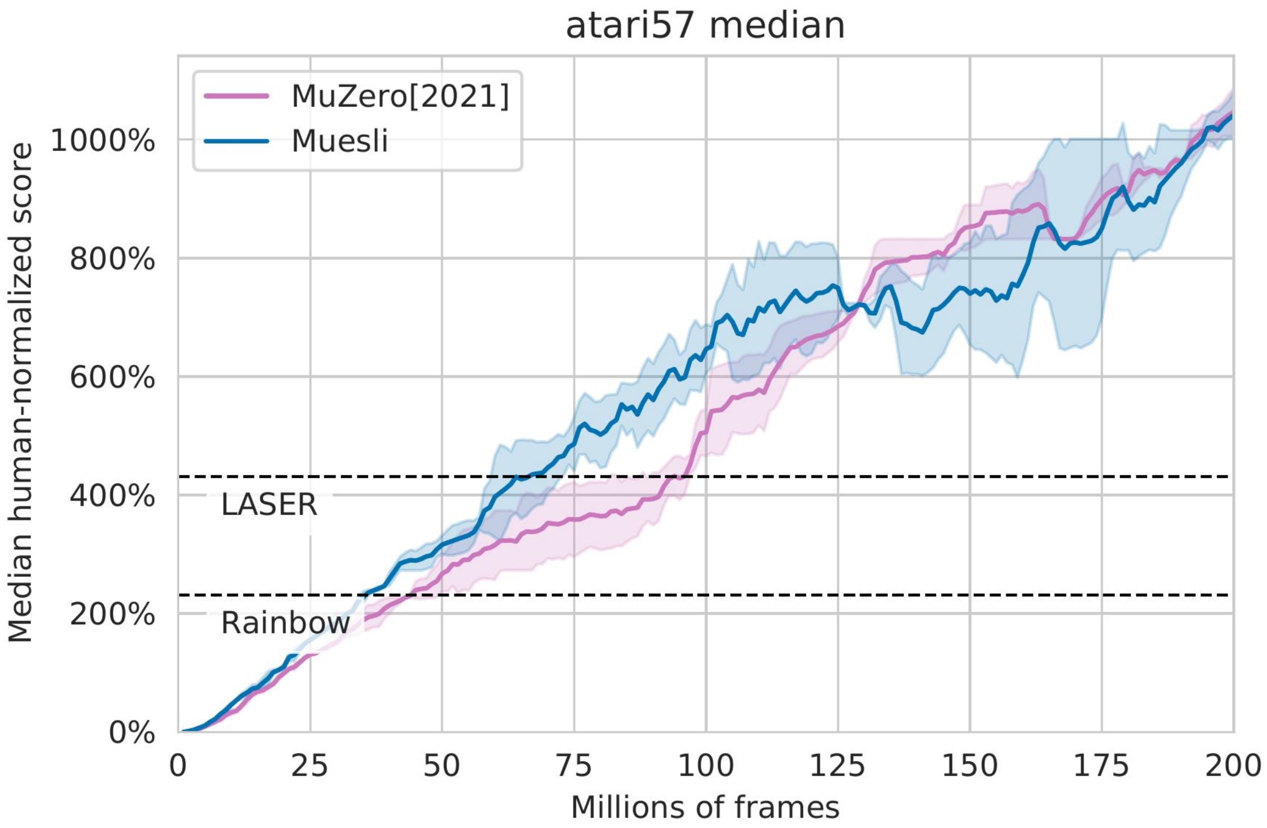
Muesli - the combined agent

Ingredients:

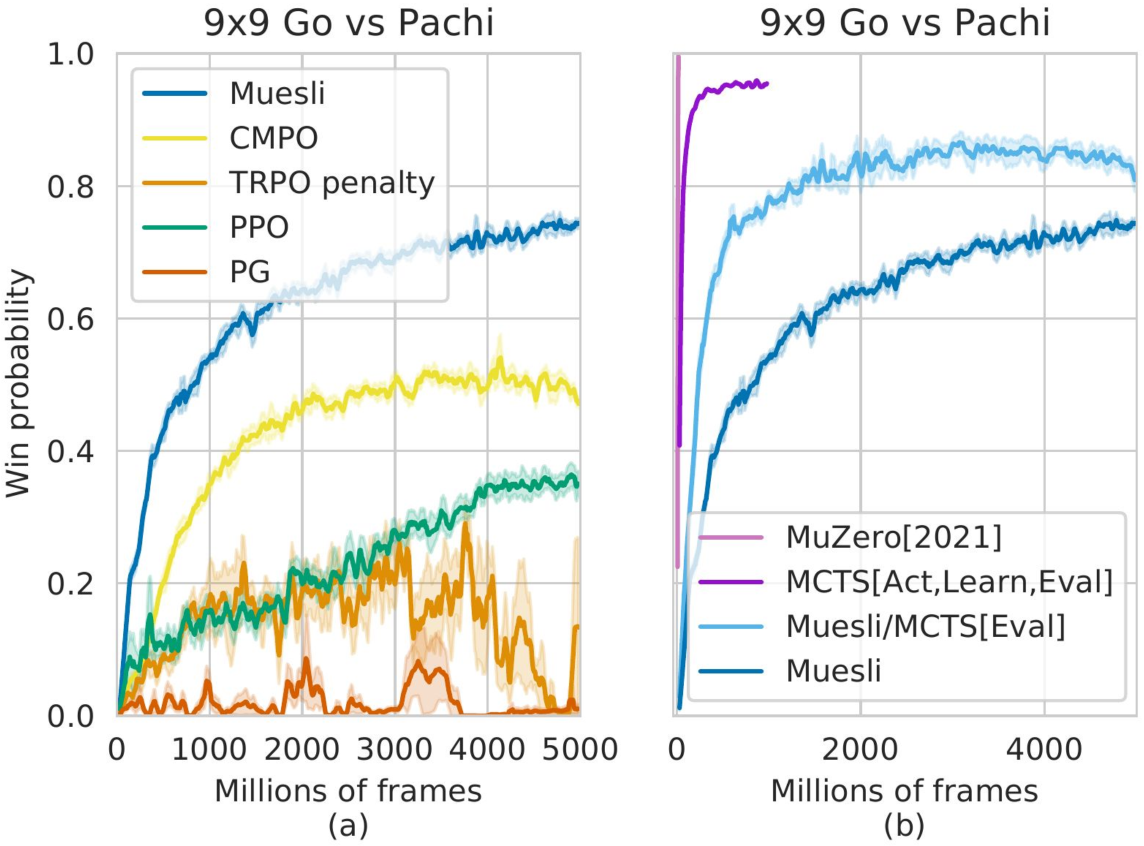
- Regularized policy optimization with Clipped MPO (CMPO).
- Retrace.
- MuZero model training as an auxiliary loss.

Acting: Directly with the policy network. No MCTS.

Atari state-of-the-art results



9x9 Go self-play results



Summary

- The value of a policy can degrade, if you compute the gradient on old data.
- The Muesli policy loss works on new environments without tuning.