Improving Stability in Deep Reinforcement Learning with Weight Averaging

Evgenii Nikishin[1], **Pavel Izmailov**[2], **Ben Athiwaratkun**[2], **Dmitrii Podoprikhin**[1,3]
Timur Garipov[4], **Pavel Shvechikov**[1], **Dmitry Vetrov**[1,3], **Andrew Gordon Wilson**[2]
[1]National Research University Higher School of Economics, [2]Cornell University
[3]Samsung-HSE Laboratory, [4]Samsung AI Center in Moscow

p(B|A)yesgroup.ru

SAMSUNG

Outline

- Deep reinforcement learning (RL) methods are notoriously unstable during training.
- Stochastic weight averaging (SWA) is a technique based on averaging the weights collected during training with an SGD-like method.
- We propose to apply SWA, in order to reduce the effect of noise on training.
- We show that SWA stabilizes the model solutions and improves the average rewards.

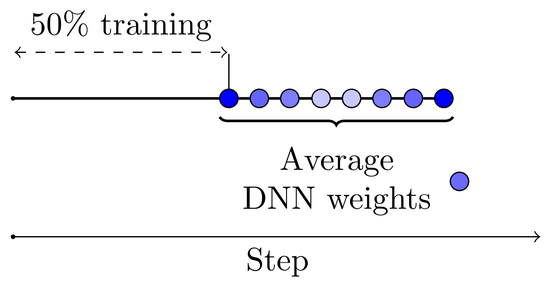
Methods

Reinforcement learning

- Advantage Actor-Critic (A2C) is a standard RL algorithm, often applied to problems with discrete action spaces.
- Deep Deterministic Policy Gradient (DDPG) is another standard RL algorithm, but suitable for continuous action spaces.

Stochastic weight averaging

SWA was shown to find solutions with better generalization in both supervised and semi-supervised learning.



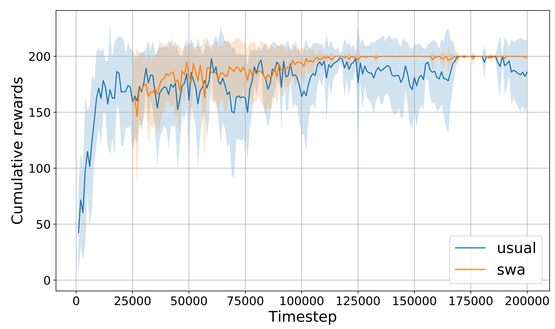
After initial pretraining phase (for example, 50% of computational budget), SWA collects weights every c timesteps.

Note that there is no need to store all the weights in memory:

$$w_{\text{SWA}} \leftarrow \frac{n_{\text{SWA}} \cdot w_{\text{SWA}} + w}{n_{\text{SWA}} + 1}, \qquad n_{\text{SWA}} \leftarrow n_{\text{SWA}} + 1$$

Results

Average cumulative rewards of A2C for CartPole environment with and without SWA



A2C on Atari environments

ENV NAME	A2C	A2C + SWA
Breakout	522 ± 34	$\mathbf{703 \pm 60}$
Qbert	18777 ± 778	$\mathbf{21272 \pm 655}$
SpaceInvaders	7727 ± 1121	$\mathbf{21676 \pm 8897}$
Seaquest	1779 ± 4	$\mathbf{1795 \pm 4}$
CrazyClimber	$\mathbf{147030 \pm 10239}$	139752 ± 11618
BeamRider	9999 ± 402	$\mathbf{11321 \pm 1065}$

DDPG on MuJoCo environments

ENV NAME	DDPG	DDPG + SWA
Hopper	613 ± 683	$\mathbf{1615 \pm 1143}$
Walker2d	1803 ± 96	$\mathbf{2457 \pm 241}$
Half-Cheetah	3825 ± 1187	$\mathbf{4228 \pm 1117}$
Ant	865 ± 899	$\mathbf{1051 \pm 696}$

Discussion

- Currently, the averaging does not affect the training procedure. Modification of the training procedure based on weight averaging can potentially help in stabilization and training acceleration.
- Theoretical justification of weight averaging in RL context.
- Analysis of the RL loss surface can reveal new approaches to weight averaging for further stabilization.