Is Policy Learning Overrated?: Width-Based Planning and Active Learning for Atari

Benjamin Ayton
MIT aytonb@mit.edu

Masataro Asai
MIT-IBM Watson AI Lab masataro.asai@mit.edu

Work supported by IBM-Watson AI Laboratory and Woodside

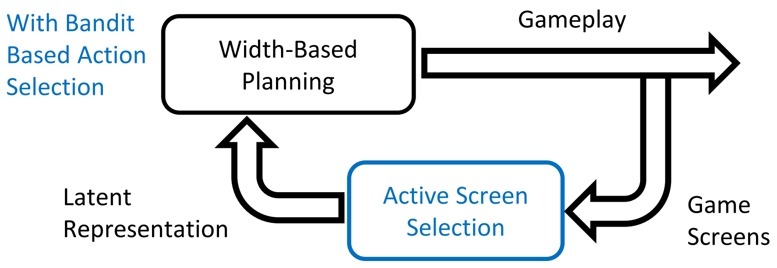
Motivation

Policy learning in Atari games results in high gameplay performance but requires 10^7+ interactions with the environment. To develop decision making suitable for novel environments with lower data requirements, we introduce **Olive**, based on **width-based planning** methods and incorporating **online and active learning.**

Overview

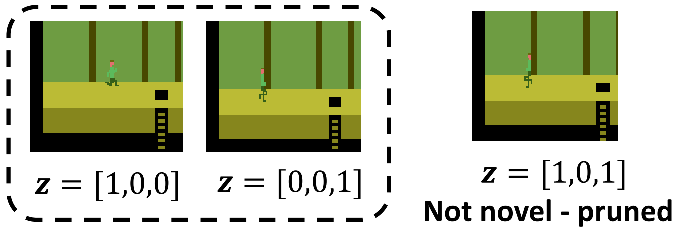
Olive performs online learning without the full complexity of policy learning. We use width based planning, making use of a latent screen representation derived from a variational autoencoder [1]. Our innovations are:

- **Active screen collection**
- **Bandit-based action selection**



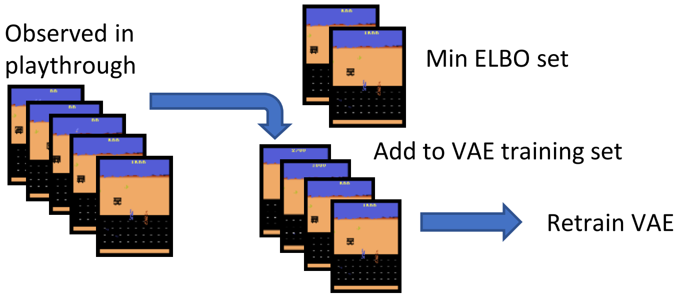
Active Screen Collection

A variational autoencoder (VAE) determines features from screens. During gameplay, screens are pruned from a search tree if they do not have a **novel feature** [2].



The VAE can only assign meaningful features for screens similar to its dataset. To capture features for later levels only reached by a performant agent, we **train the VAE online** through multiple playthroughs.

To train an effective VAE with limited data, we apply **uncertainty sampling** and select **screens with minimum evidence lower bound** (ELBO).



Bandit Based Action Selection

Rewards of each action are modeled with a normal inverse chi-squared prior. During planning, actions are explored according to **Top Two Thompson Sampling (TTTS)** [3], based on **best-arm identification** problems in multi-armed bandits. TTTS assists in finding high reward actions more rapidly.

Results

Olive with active screen selection (ActiveOlive) outperforms width based planning with a VAE trained offline (VAE-IW) [1].

ActiveOlive 32 wins – 20 wins VAE-IW

ActiveOlive (10^5 interactions) outperforms policy learning methods π-IW [4] and deep Q networks (DQN, 10^7+ interactions) [5], and EfficientZero, (10^5 interactions) [6].

ActiveOlive 30 wins – 22 wins π-IW

ActiveOlive 31 wins – 17 wins DQN

ActiveOlive 18 wins – 7 wins EfficientZero

[1] Dittadi, A.; Drachmann, F. K.; and Bolander, T. 2021. Planning from Pixels in Atari with Learned Symbolic Representations. In AAAI, volume 35, 4941–4949.
[2] Lipovetzky, N.; and Geffner, H. 2012. Width and Serialization of Classical Planning Problems. In ECAI, 540–545.
[3] Russo, D. 2020. Simple Bayesian Algorithms for Best-Arm Identification. Operations Research, 68(6): 1625–1647.
[4] Junyent, M.; Jonsson, A.; and Gomez, V. 2019. Deep Policies for Width-Based Planning in Pixel Domains. In ICAPS, volume 29, 646–654.
[5] Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-Level Control through Deep Reinforcement Learning. Nature, 518(7540): 529–533.
[6] Ye, W.; Liu, S.; Kurutach, T.; Abbeel, P.; and Gao, Y. 2021. Mas tering atari games with limited data. Neurips, 34.