Training long history on real reward and diverse hyper parameters in threads combined with DeepMind's A3C+

Takayoshi Iitsuka
The Whole Brain Architecture Initiative, Japan

Games with little chance of scoring such as Montezuma's revenge are difficult for Deep Reinforcement Learning (DRL) because there is little chance to train Neural Network (NN), i.e. no reward, no learning. DeepMind indicated that pseudo-count based pseudo-reward is effective for learning of games with little chance of scoring. They achieved over 3000 points in Montezuma's revenge by combination with Double-DQN. On contrary, its average score was only 273.70 point in combination with A3C (it is called A3C+). A3C is very fast training method and getting high score with A3C+ is important. I propose new training methods: Training Long History on Real Reward (TLHoRR) and Diverse Hyper Parameters in Threads (DHPT) for combination with A3C+. TLHoRR trains NN with long history just before getting score only when game environment returns real reward i.e. training length by real reward is over 10 times longer than that of pseudo-reward. This is inspired by reinforcement of learning with dopamine in human brain. In this case, real score is very valuable reward in brain and TLHoRR strongly trains NN like dopamine does. DHPT changes hyper parameters of learning in each thread and make diversity in threads actions. DHPT was very effective for stability of training by A3C+. Without DHPT, average score is not recovered from zero when it is dropped to zero. With TLHoRR and DHPT in combination with A3C+, average score of Montezuma's revenge almost reached 2000 points. This combination made exploration of game state better than that of DeepMinds's paper. In Montezuma's revenge, five rooms are newly visited by TLHoRR and DHPT; they were not visited by DeepMinds's pseudo-count based pseudo-reward only. Furthermore, with TLHoRR and DHPT in combination with A3C+, I got and kept top position in Montezuma's revenge in OpenAI gym environment from October 2016 to March 2017.

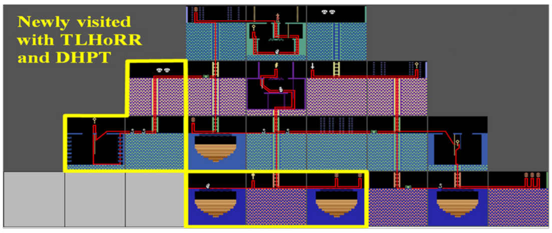

Figure 1: Effects of Training Long History on Real Reward (TLHoRR) and Diverse Hyper Parameters in Threads (DHPT) . Rooms surrounded by yellow line are newly visited by TLHoRR and DHPT, not visited by DeepMind's pseudo-count based pseudo-reward only.

Biography

Takayoshi Iitsuka completed his Master's degree in Science and Technology at University of Tsukuba in Japan. From 1983 to 2003, he was a Researcher and Manager of optimizing and parallelizing compiler for supercomputers in Central Research Laboratory and Systems Development Laboratory of Hitachi. From 2003 to 2015, he was in strategy and planning department of several IT divisions. He retired from Hitachi in October 2015 and started study and research of Artificial Intelligence in May 2016. In October, he achieved top position of Montezuma's revenge in OpenAI gym. His current research interests include Deep Learning, Deep Reinforcement Learning and Artificial General Intelligence based on whole brain architecture.

iitt21-t@yahoo.co.jp

Notes: