Policy Distillation

Qilei Zhang

Jun 20 2018

Abstract

Policies for complex visual tasks have been successfully learned with deep reinforcement learning, using an approach called deep Q-networks (DQN), but relatively large (task-specific) networks and extensive training are needed to achieve good performance.

1. Introduction

Recently, advances in deep reinforcement learning have shown that policies can be encoded through end-to-end learning from reward signals [1], and that these pixel-to-action policies can deliver superhuman performance on many challenging tasks [2].

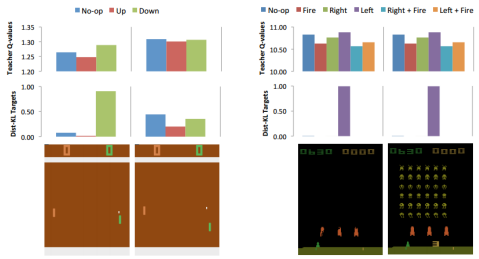

Figure 1. Example frames from two Atari games, with the Q-values output by DQN (top) and the distillation targets after softmax (middle). For Pong, the two frames only differ by a few pixels yet the Q-values are different. In the Space Invaders example, the input frames are very different yet the Q-values are very similar. In both games the softmax sharpens the targets, making it easier for the student to learn.

$$x'(t) = -V'(x) + A_0 cos(wt + o) + u(t) \qquad (1)$$

2. Previous Work

This work is related to four different research areas: model compression using distillation, deep reinforcement learning [3], multi-task learning and imitation learning. The concept of model compression through training a student network using the outputs of a teacher network was first suggested by Bucila who proposed it as a means of compressing a large ensemble model into a single network [4].

	DQN	Dist-MSE	Dist-NLL	Dist-KL
	score	score	score	score
Breakout	303.9	102.9	235.9	287.8
Freeway	25.8	25.7	26.2	26.7
Pong	16.2	15.3	15.4	16.3
Qbret	4589.8	5607.3	6773.5	7112.8

Table 1. Comparison of learning criteria used for policy distillation from DQN teachers to students with identical network architectures: MSE (mean squared error), NLL (negative log likelihood), and KL (Kullback-Leibler divergence). Best relative scores are outlined in bold

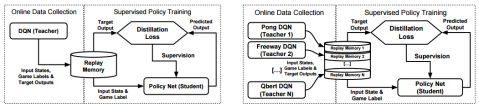

Figure 2. (a) Single-task data collection and policy distillation. The DQN agent periodically adds gameplay to the replay memory while the student network is trained. (b) Multi-task data collection and policy distillation.

3. Approach

Before describing policy distillation, we will first give a brief review of deep Q-learning, since DQN serves as both the baseline for performance comparisons as well as the teacher for the policy distillation. Note that the proposed method is not tied to DQN and can be applied to models trained using other RL algorithms. After the DQN summary, we will describe policy distillation for single and multiple tasks [5].

References

[1] Gaminibandara and G. K. K. G. Synthesis of optimal distillation systems. *Contemporary Security Policy*, 20(3):198–230, 1976. 1

[2] Kupferberg and Arie. The mechanism of dumping sieve trays in distillation columns. *Energy Policy*, 64(5):203C208, 1968. 1

[3] P. Li and G. Wozny. Adaptive control of multiple-fraction batch distillation for tracking optimal a priori policies. *Inzynieria Chemiczna I Procesowa*, 19(3):511–529, 1998. 1

[4] G. Wozny and P. Li. Optimisation and experimental verification of startup policies for distillation columns. *Computer Aided Chemical Engineering*, 28(1):253–265, 2002. 1

[5] Z. Xue-mei, J. Chun-gui, R. Hong-dong, and Z. Wei-jiang. A novel operation policy for slop cut of batch distillation. *Chemical Engineering*, 34(10):5–8, 2006. 1