AUTOMATIC RL CURRICULUM VIA PECULIAR STATES IDENTIFICATION

Jakub Lewkowicz, Mikhail Lanchytski, Bartłomiej Kocot, Paweł Czarnul, Piotr Januszewski

Gdańsk University of Technology

Problem statement

The Google Research Football provides many training academies such as typical football situations (corner, counter-attack, etc.) but also basic tasks such as scoring a goal. Those academies might be used to create a learning scenario, but would be this curriculum optimal?

- Robust evaluation process.
- 11 football academies might be not enough to create a curriculum.

Such a process seems robust but does not guarantee success. To minimize the cumbersome process of creating a curriculum, our team proposes a novel method.

- Might be used in a different OpenAI Gym environments.
- Easy to evaluate.
- Generates adaptive curriculum.

Finding interesting states

TD-error(n) is a quantity measuring error in estimates over the next n steps. TD-error(n) $= V(S_k) + \sum_{t=k+1}^{k+n} R_{t+1} - V(S_t)$, starting from the state S_k. TD-error might be used to indicate if a given state is interesting or not. It may estimate if the agent was surprised or if the value function overestimated the state's value.

- TD-errors are calculated over n steps.
- TD-error over 1 step is not enough to find interesting state, 5 steps seem optimal.

To find the most interesting states TD-errors have to be calculated for all episode states and then only peak values shall be selected.

- Small TD-error –> small mistake of a value function.
- Abnormal TD-error –> peculiar state.

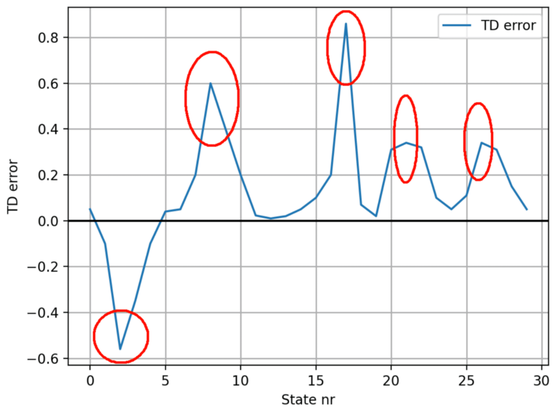

Fig. 1: TD-error peak values

- TD-errors in the surrounding of the peak value lead to the same interesting situation.
- Both local minimums and maximums may indicate the peculiar state

Generation of adaptive training scenarios for curriculum learning

Our method is based on finding states in the agent experience for which the agent couldn't properly infer their value. To achieve it, we utilize the TD-error of the actor-critic agent's value function. States with high TD-error are considered interesting to the agent and stored in a buffer. Interesting situations are restored alternately with resetting the environment in further training.

- Gives the agent the possibility to practice situations in which its value function had a high error.
- Different than an off-policy replay buffer in a way that the agent has a chance to generate new experience in the situation it found difficult.
- Ease of usage and adaptability of the method to different Gym environments.
- Generation of a suited curriculum.
- Improves overall agent's performance.

Training with peculiar states

The following algorithm presents how to adapt our method to PPO-Clip training.

Algorithm 1: PPO-Clip + automatic curriculum (highlited)

Input: initial value function parameters ϕ_0, initial policy parameters θ_0 and frequency parameter $freq$.
for $k = 0, 1, 2, ...$ **do**
  Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
  Compute rewards-to-go $\hat{R}_t$.
  Compute TD-errors.
  Find interesting states by selecting peak values of computed TD-errors.
  Update the buffer.
  Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function V_{ϕ_k}.
  Update the policy by maximizing the PPO-Clip objective.
  Fit value function by regression on mean-squared error.
  if $k \mod freq == 0$ **then**
    Start the new episode with the randomly chosen interesting state from the buffer.
  else
    Reset the environment.
  end if
end for

Google Research Football (GRF)

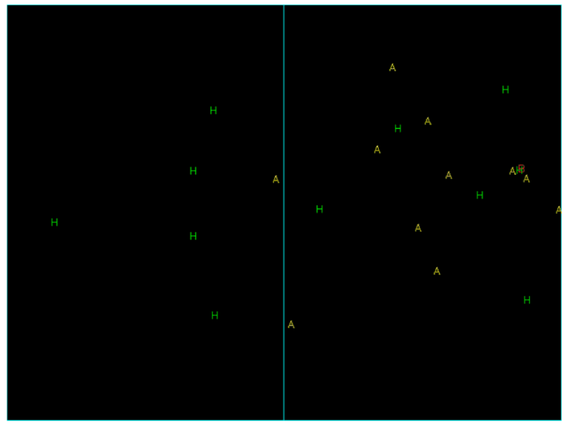

Fig. 2: H: agent's players. A: enemy's players. B: ball.

- Physics-based 3D football simulator.
- 11 different football tasks.
- Open-source.

Fig. 2 presents an interesting state found during the training with our method.

- Agent's player in the goal area.
- Agent is ready to take a shot.
- Opponents close to the agent's player are trying to recapture the ball.

At one of the next steps, our agent scored a goal. Despite winning the game agent struggled to score a goal, therefore the state was labeled as peculiar.

Results

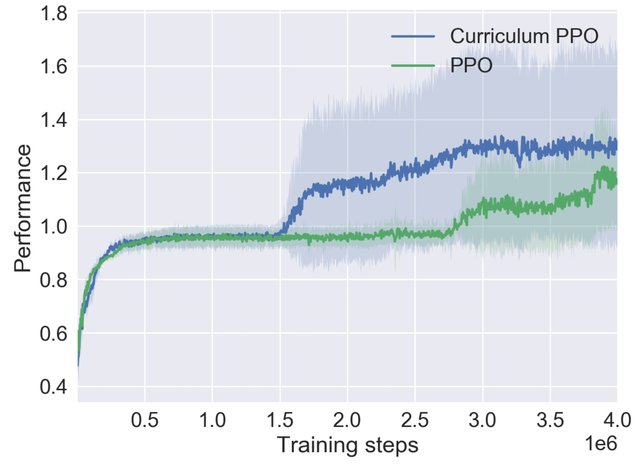

Fig. 3: Methods comparison on "Pass and Shoot with Keeper" football academy. Episode return median across 5 seeds.