Deep RL from Human Preferences

June 11, 2022

Problem

Designing reward function for complex RL system is hard

Approaches

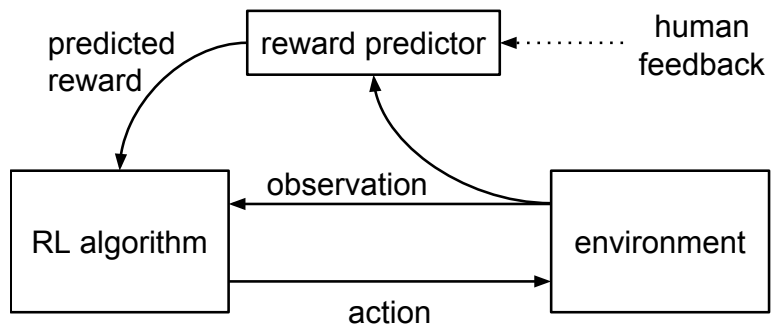
- **Learn it** Learn what is good behavior vs bad behavior through supervised learning
- **Inverse RL** Given trajectories from an optimal policy, infer a reward function that results the optimal policy. Ng, Russell 2000
  - many reward functions can result in the same optimal policy

The paper explores approach 1.

Overview of the methods

- collect RL agent's behavior
- send video clips of those behavior to human labelers
- human labelers see two behaviors at once, they provide feedbacks on which one of them is better
- train a reward predictor to fit human preference
- continue training the RL agent to please the reward predictor

See section
2 of the paper for more details



Problems with the methods

- cost of labels
- reward predictor might return a wrong scale
- when the RL policy is near random, comparison tasks for human labelers are not meaningful

Design space

- Single reward predictor or an ensemble of reward predictors? They choose ensemble as baseline.
- When to query for human feedbacks? They use variance among predictors as a heuristic measure of information gain.
- how long the video segments to show to human lablers. They used vidoe segmentas lasted from 3 - 5 seconds.

See Ablation Studies in 3.3 for a complete list of design choices.

Experiment results

Experiements performed on MuJoCo and Atari. They compared RL agent's performance when trained with

- traditional reward
- synthetic feedbacks from an oracle. The oracle has access to true reward, and provide feedbacks based on the true reward
- human feedbacks

High level observations:

- real reward does not result the best performance in many cases
- human feedback on Ant resulted much faster learning

See section 3 for more details

Experiment results (continued)

Learning novel behaviors. See section 3.2

Experiment results (continued)

Why human feedbacks are especially useful for Ant?
Link to demo
Reward:

$$r_t = x_t - 0.5 \times ||a_t||_2^2 - 0.0005 \times ||s_t||_2^2 + 1$$

Link to MuJoCo env reward

Experiment results (continued)

Why Qbert has such a bad performance when trained with human feedbacks
Link to Demo
My guess: 3-5 vidoe clips is not sufficient for human lablers to figure out what is a good behavior.

Ablation Studies

See section 3.3 for details.

- impactful factors: length of segment, ensemble
- counter-intuitive observations: uncertainty based query strategy does not seem to outperform random queries