Human Influence for Reinforcement Learning

Deep Q-learning from Demonstrations
T. Hester et al.

Deep Reinforcement Learning from Human Preferences
P. Christiano et al.
Conventional Reinforcement Learning

Learning from Demonstrations

- Imitation learning:
  - By design never outperform human experts
  - Only exploit narrow area of state-action space

- Combined reinforcement and imitation learning:
  - Reward / policy shaping

- Teacher / apprenticeship agents:
  - Learning from trained agents
Deep Q-learning from Demonstrations
Deep Q-learning from Demonstrations
Deep Q-learning from Demonstrations

- Expert Controller
- Storage
- DRL Controller
- System
Base Network

- Double DQN with prioritized experience replay [1,2]
  - Double DQN: reduced reward overestimation
  - Prioritized experience replay: increased number of hard tasks

\[ J_{DQ}(Q) = (R(s, a) + \gamma Q(s_{t+1}, a_{t+1}^{\max}; \theta') - Q(s, a; \theta))^2 \]


Two Phase Learning

- **Pre-training (offline)**
  - Replay buffer:
    - Controller data
  - Loss:
    - 1-step double Q-learning loss
    - n-step double Q-learning loss (n=10)
    - Supervised large margin classification loss
    - L2 regularization loss

- **Online learning**
  - Replay buffer:
    - Controller data (not overwritten + prioritized)
    - Self-generated data
  - Loss:
    - 1-step double Q-learning loss
    - N-step double Q-learning loss
    - (Supervised large margin classification loss) for controller data
    - L2 regularization loss
Loss Function

- Supervised large margin classification loss [1]
  - Limits value of unseen actions
    \[ J_E(Q) = \max_{a \in A} [Q(s, a) + l(a_E, a)] - Q(s, a_E), \quad l(a_E, a) = \begin{cases} 
      0 & a = a_E \\
      c & a \neq a_E 
    \end{cases} \]

- 1-step + N-step double Q-learning loss
  - Guarantee Bellman equation

- L2 regularization loss
  - Network weight + bias regularization

Experiments

- 42 Atari games played 3-12 times ➔ 5,574 to 75,472 transitions/game
  - Outperforms worst demonstration in 29 games
  - Outperforms best demonstration in 14 games
Results

Related Work: Montezuma Revenge

- ADET
- DQfD
- Human Experience Replay
- Replay Buffer Spiking

Related Work: Qbert

- ADET
- DQfD
- Human Experience Replay
- Replay Buffer Spiking
Results

Ablation Study

Loss Ablations: Montezuma Revenge

- Blue line: DQfD
- Black dashed line: No Supervised Loss
- Green dotted line: No n-step TD loss

Training Episode Returns vs. Training Iteration
Demonstration Up-Sample Ratio

![Graph showing demonstration data up-sample ratio over training iterations for different games such as Hero, Montezuma's Revenge, Pitfall, Q-Bert, and Road Runner.](image-url)
We can intuitively define complex reward functions!

Deep Reinforcement Learning from Human Preferences

- Solve DRL tasks without observing the true reward
- Comparison of video sequences ➔ intuitive evaluation
- Not contingent on human performing task
- Potential to outperform conventional DRL
Deep Reinforcement Learning from Human Preferences

- **DRL Controller**
- **Reward Predictor**
- **Environment**
- **Observation**
- **Action**
- **Human Feedback**
- **Predicted Reward**
Deep Reinforcement Learning from Human Preferences

- Conventional DRL step:
  \[(o_i, a_i) \rightarrow r_i\]

- Trajectory segment:
  \[\sigma = ((o_0, a_0), (o_1, a_1), ..., (o_{k-1}, a_{k-1})) \rightarrow r_{k-1}\]

- Human can rate order of trajectory segments:
  - Goal in human language
  - Present video segments of agent’s attempts
  - Rate videos \(\sigma^1 \succ \sigma^2\)
Training pipeline

- **DRL with predicted rewards:**
  - Interaction with environment
  - Trajectory generation

- **Human evaluation:**
  - Trajectory comparison

- **Reward predictor training:**
  - Optimization of reward predictor
DRL with Predicted Rewards

- **Tasks:**
  - Interaction with environment
  - Generation of trajectories

- **Methods:**
  - Conventional DRL with non-stationary reward function
  - Atari: advantage actor critic (A2C) [1]
  - Robots: trust region policy optimization (TRPO) [2]


Human Evaluation

- 1s – 2s segments are evaluated

- Database $\mathcal{D}$ of triples $(\sigma^1, \sigma^2, \mu)$

- Queries based on prediction variance $\Rightarrow$ approximates value of information
Reward Predictor Training

- Preference predictor: latent factor of human judgement
  \[
  \hat{P}[\sigma^1 > \sigma^2] = \frac{\exp \sum \hat{r}(o^1_t, a^1_t)}{\exp \sum \hat{r}(o^1_t, a^1_t) + \exp \sum \hat{r}(o^2_t, a^2_t)}
  \]

- Training with cross entropy loss
  \[
  \text{loss}(\hat{r}) = -\sum_{(\sigma^1, \sigma^2, \mu) \in D} \mu(1) \log \hat{P}[\sigma^1 > \sigma^2] + \mu(2) \log \hat{P}[\sigma^2 > \sigma^1]
  \]

- Implementation details:
  - Ensemble of predictors
  - L2 regularization optimized on validation set
  - Assumption: Human choice 10% at random
Results Simulated Robotics

Results Atari

Complex Task: Hopper Backflip

Complex Task: Half-Cheetah Handstand

Complex Task: Enduro keep alongside cars

Ablation Simulated Robotics

Ablation Results

- Offline reward predictor training results in strange behavior
- Querying comparisons is more helpful than absolute scores
- Sequences are more helpful than single frames
Summary

- DRL for hard tasks can profit from human intuition
- Boost initial performance with demonstrations
- Behavioral ratings for not directly solvable tasks