Curiosity-driven Reinforcement Learning for Dialogue Management

Paula Wesselmann, Yen-Chen Wu, Milica Gašić

Introduction
- Dialogue Manager (DM) is the brain of the dialogue system
- DM tracks believes and determines behaviour of system
- DM uses Reinforcement Learning (RL) to learn a policy
- RL is learning from feedback/rewards

Motivation
- Hard to obtain user feedback/external reward
- Explore more efficiently
- Improve policy learning

Policy learning \( \pi(b) : B \rightarrow A \)
- Choose policy that maximises total Reward
- \( Q(b, a) = \sum_{k=0}^{T-i} \gamma^k r_{i+k} | b_i = b, a_i = a \)
- \( \pi^*(b) = \arg \max_a Q^*(b, a) \)
- Q-function represented by deep Neural Network
- Policy optimisation with DQN
- For DQN policy is used eps-greedily to determine action

Intrinsic Reward Signal
- RL relies on reward signals (usually external feedback)
- For Dialogue systems those reward signals are often hard to obtain, not accurate or even absent
- Intrinsic reward systems such as curiosity, can replace external feedback or be used in addition to external rewards
- Explore more efficiently by actively seeking new knowledge, no random exploring

Intrinsic Curiosity Module (ICM)
- State prediction error as curiosity reward (Pathak et al. 2017)
- No random exploration needed anymore i.e. no eps-greedy

Handcoded Curiosity Experiments
- Increased initial exploration (random)
- Vary the use of turn penalty as reward signal

Preliminary Results
- Most simple environment, only one seed;

Next Steps
- Tuning the reward signal and other parameters
- Intrinsic reward signals only
- Predicting using larger (more specific) action space
- Alternative curiosity rewards to state prediction error
- Implement in hierarchical framework

Reference:
Pathak, D. and Agrawal, P. and Efros, A. A. and Darrell, T.

PyDial
- CUED Python Statistical Dialogue System

The Dialogue Manager:
1. Updates belief state of system \( b_i \)
2. Selects action \( a_i \)

Intrinsic Curiosity Module (ICM)
- State prediction error as curiosity reward (Pathak et al. 2017)
- No random exploration needed anymore i.e. no eps-greedy

Introduction
- Policy optimisation with DQN
- For DQN policy is used eps-greedily to determine action

Motivation
- Hard to obtain user feedback/external reward
- Explore more efficiently
- Improve policy learning

Policy learning \( \pi(b) : B \rightarrow A \)
- Choose policy that maximises total Reward
- \( Q(b, a) = \sum_{k=0}^{T-i} \gamma^k r_{i+k} | b_i = b, a_i = a \)
- \( \pi^*(b) = \arg \max_a Q^*(b, a) \)
- Q-function represented by deep Neural Network
- Policy optimisation with DQN
- For DQN policy is used eps-greedily to determine action

Intrinsic Reward Signal
- RL relies on reward signals (usually external feedback)
- For Dialogue systems those reward signals are often hard to obtain, not accurate or even absent
- Intrinsic reward systems such as curiosity, can replace external feedback or be used in addition to external rewards
- Explore more efficiently by actively seeking new knowledge, no random exploring

Intrinsic Curiosity Module (ICM)
- State prediction error as curiosity reward (Pathak et al. 2017)
- No random exploration needed anymore i.e. no eps-greedy

Handcoded Curiosity Experiments
- Increased initial exploration (random)
- Vary the use of turn penalty as reward signal

Preliminary Results
- Most simple environment, only one seed;

Next Steps
- Tuning the reward signal and other parameters
- Intrinsic reward signals only
- Predicting using larger (more specific) action space
- Alternative curiosity rewards to state prediction error
- Implement in hierarchical framework

Reference:
Pathak, D. and Agrawal, P. and Efros, A. A. and Darrell, T.