Adversarial Advantage Actor-Critic Model for Task-Completion Dialogue Policy Learning

Baolin Peng, Xiujuan Li, Jianfeng Gao, Jingjing Liu, Yun-Nung (Vivian) Chen, Kam-Fai Wong

Summary

- **Advantage Actor-Critic for Dialogue Policy Learning**
  - Find a policy π that maximizes the expected reward: $R = \sum_{t=0}^{T-1} \gamma^t r_t$
  - A parameterized probabilistic mapping function: $\tau_\theta(a|s) = P(A_t = a | s_t = s; \theta)$
  - Update $\theta$ with following gradient: $\nabla_\theta J(\theta) = E[\nabla \log \tau_\theta(a|s) Q(s,a,\theta) - Q(s,a)]$
  - Baseline function for reducing variance: $E[\nabla \log \tau_\theta(a|s) Q^r(s,a,\theta)] = E[Q^r(s,a) - V^r(s)]$
  - TD error as an unbiased estimation: $\nabla_\theta J(\theta) = E[\nabla \log \tau_\theta(a|s) Q^r(s,a,\theta)] - \nabla \log \tau_\theta(a|s) V^r(s)$

- **Adversarial Training**
  - Actor π as a generator G
  - A discriminator D identifies state-action pair (s, a) from experts or G
  - D can be viewed as a reward function extracted from experts’ trajectories
  - D is to maximize the probability of classifying each pair correctly: $\log D(s, a; \theta_D) = \log (1 - D(s, a; \theta_D))$
  - Actor $\pi_\theta$ (G) can be improved with $-\log (1 - D(s, a))$ as the reward function: $\nabla_\theta J(\theta) = E[\nabla \log \tau_\theta(a|s) \gamma^2 V^G_{\text{GAN}}(s) - \gamma V^G_{\text{GAN}}(s)]$
  - Combine A2C with a reward function learned from experts’ demonstrations with adversarial training.
  - The discriminator D guides actor to explore state action regions where human experts will explore.

2. Methodology

- **Discriminator Training**
  - Expert Demonstration
  - Sample (s,a) pair
  - Discriminator
  - Simulation
  - Sample (s,a) pair
  - Discriminator
  - Actor
  - TD error
  - Reward
  - User Simulator
  - Actor
  - TD error
  - Critic
  - Discriminator
  - System Action
  - a
  - Adversarial Advantage Actor-Critic

3. Experiments & Results

- **Dataset**: human-human conversations in the movie-ticket booking scenario
  - collected via AMT and annotated by human experts
  - 280 labeled dialogue with 11 average turns
  - 11 dialogue acts, 29 slots
  - Informable (narrow down search), requestable (ask info from agent)
  - Use a publicly available user simulator

- **Baselines**
  - Rule Agent: handcrafted rule-based policy that in- forms and requests a hand-picked subset of necessary slots.
  - A2C trained with a pre-defined reward function and a standard advantage actor-critic algorithm
  - BBQN-Map Agent (AAAI’18): the best agent among a set of BBQN variants that has great efficiency for policy exploration in dialogue systems

- **Evaluation**
  - Success rate
  - 10 run averaged learning curve
  - 2000 dialogues for testing

- **Adversarial A2C learns faster and more stable with better exploration.**

4. Conclusion

- **We propose an adversarial advantage actor-critic model with efficient exploration.**
  - The discriminator serves as an **additional critic** to guide policy exploration towards human-like one.
  - It also has connection with inverse reinforcement learning that learns reward function.
  - Our experiments in a movie-ticket booking domain show its superiority.