RLHF を用いた人間に寄り添うエージェントの研究

横山　凜　【 強化学習＆探索研究室 】

1　はじめに

強化学習において，意図した挙動を誘導する報酬設計は困難であり，人間の選好に基づく RLHF[1] が有力視されている．本手法の挙動特性を正確に把握するためには，手動での報酬設計が容易であり，最適行動が自明な環境での基礎的な検証が有用である．本稿では，目的とする行動が明確であり，実装が容易であるグリッドワールド環境を用い，本手法が意図した行動を正確に再現できるか検証した結果を報告する．

2　検証手法

図 1 に検証手法の概要を示す．(1) 選好データセットの構築，(2) 報酬モデルの学習，(3) 学習された報酬を用いたエージェントの訓練，の 3 段階で構成される．

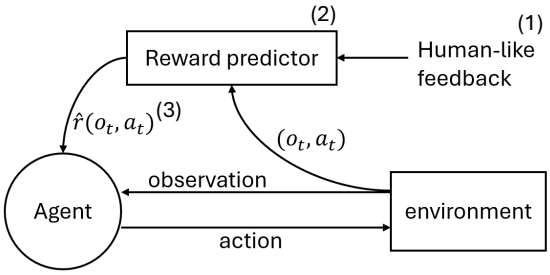

図 1　検証手法の概要

2.1　選好データセットの構築

本研究では学習済みモデルを用いた擬似フィードバック（Human-like Feedback）を導入した．具体的には，異なる行動を取る複数のエージェントを用意し，それらの軌跡に対して「人間らしさ」と「ゴール到達」の両面から主観的にランク付けを行う．例えば，「コインを獲得してからゴールに向かう軌跡」を最高ランク，「コインを無視してゴールへ直行する軌跡」を低ランクとしてラベル付けする．これにより，ゴール到達という目標情報と行動の選好をとらせる目標情報が選好ラベルに暗黙的に埋め込まれる．

2.2　報酬モデルの学習

報酬予測器 $R_{learned}$ には，観測を平坦化して入力する全結合ニューラルネットワーク（MLP）を採用している．学習には，Christiano et al.[1] に基づく Bradley-Terry モデルの損失関数（式 1）を用いる．y は $\sigma^1 \succ \sigma^2$ を表す選好ラベルである：

$$\mathcal{L} = -\sum_{(\sigma^1,\sigma^2,y)} \log \frac{\exp \sum \hat{r}(\sigma^1)}{\exp \sum \hat{r}(\sigma^1) + \exp \sum \hat{r}(\sigma^2)} \quad (1)$$

2.3　強化学習によるポリシー獲得

学習済み $R_{learned}$ を用いて PPO によりエージェントを訓練する．R_{env} を使用せず，選好データから学習された報酬のみでポリシーを更新する．ゴール到達の情報は選好ラベル構築時に反映されているため，$R_{learned}$ は進捗度に応じた密な報酬信号を出力する．

3　実験設定

MiniGrid 上のグリッドワールド環境において，エージェントと人間の選好が異なる複数の環境を用意した．一例として，「コインの獲得」という隠されたタスクが存在する環境（図 2）で「コインを獲得してからゴールへ向かう」軌跡を高ランクとした選好データを作成し，R_{env} を用いない状況下において，選好学習のみから意図した行動を正確に再現できるか検証した．

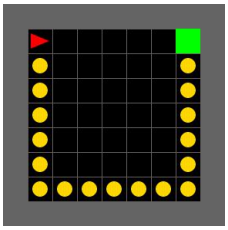

図 2　実験環境の一例（MiniGrid-Coin）

4　実験結果と考察

実験の結果，明示的報酬を与えた PPO と同等のステップ数で学習が完了し，選好ラベルのみ意図した行動構造を学習できた．これは，報酬予測器が密な報酬信号を出力することで明示的なサブゴール報酬が利用できない状況下でも効率的な探索が可能となったためである．以上より適切な選好データセットにより明示的な報酬設計なしに意図した行動構造を学習可能であることが示された．

5　まとめ

本稿では，選好ラベルにゴール到達目標と望ましい振る舞いへの選好目標を埋め込み，ゴール到達と人間らしい挙動を両立する手法の有効性を検証し，PPO との比較検証を報告した．

参考文献

[1] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D., “Deep Reinforcement Learning from Human Preferences”, Advances in Neural Information Processing Systems, 30, 2017.