Actor attention による協調タスクにおけるマルチエージェントの行動説明

ER19035 佐川統夢　　指導教授：藤吉弘亘

1. はじめに

マルチエージェント深層強化学習は，複数のエージェントを同一環境で学習する手法である．マルチエージェント深層強化学習の応用例には，サッカーロボットなどが挙げられる [1]．サッカーロボットはエージェント間でのパス回しなど協調行動を獲得することで，目的であるゴールを達成している．一方で，マルチエージェント深層強化学習において，獲得した協調行動に対する判断根拠が不明という問題がある．そこで本研究では，Actor-Critic 法の Actor に対し，アテンションヘッドを導入する．これにより，対象エージェントの注視度を表す Actor attention を獲得でき，協調行動に対する判断根拠の解析が可能となる．

2. Multi-Actor-Attention-Critic

Multi-Actor-Attention-Critic (MAAC) [2] とは，Actor-Critic 法の Critic にアテンションヘッドを導入したマルチエージェント深層強化学習手法である．MAAC Critic は，アテンションヘッドを用いて他エージェントの貢献度を算出する．そして，その貢献度から対象エージェントの行動価値関数を求めることで協調行動を獲得している．このとき，MAAC Critic に用いる各エージェントの状態と行動を入力とする MLP とアテンションヘッドのパラメータはエージェント間で共有している．これにより，効率的に協調行動を獲得できる．

3. 提案手法

本研究は協調行動に対する判断根拠の解析を目的とし，Actor にアテンションヘッドを導入した手法を提案する．Actor では，各エージェントの行動を選択するために，アテンションヘッドを導入し，対象エージェントの行動選択に対する他エージェントの貢献度を獲得する．提案手法の構造を図 1 に示す．

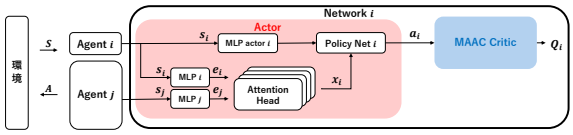

図 1：提案手法の構造

Actor の MLP は，対象エージェント i と対象エージェントを除いた他エージェント j の状態 $\mathbf{s}_i$，$\mathbf{s}_j$ を入力し，特徴量 $\mathbf{e}_i$，$\mathbf{e}_j$ を算出する．MLP で算出される特徴量 $\mathbf{e}_i$，$\mathbf{e}_j$ をアテンションヘッドに入力し，各要素間で内積し，次元数で正規化する．そして，Softmax 関数によりアテンションを算出する．また，アテンションと他エージェント j の特徴量 $\mathbf{e}_j$ から内積により他エージェント j の貢献度 $\mathbf{x}_i$ を算出する．本手法ではアテンションヘッドで算出したアテンションを Actor attention と呼称する．Actor attention は，対象エージェントがどの他エージェントを注視しているかを示す値である．Actor attention の算出式を式 (1) に示す．

$$att_{actor} = \mathrm{Softmax}\left(\mathbf{e}_i \cdot \mathbf{e}_j^T / \sqrt{dim}\right) \tag{1}$$

ここで $\mathbf{e}_j^T$ は転置した $\mathbf{e}_j$，dim は特徴量 e_i の次元数である．獲得した Actor attention を可視化することで，協調行動に対する判断根拠の解析が可能となる．

4. 評価実験

提案手法の有効性を確認するため，Multi Particle Environments の Simple_spread タスクを対象として評価する．

4.1. 実験環境

Simple_spread は各エージェントがそれぞれランドマークに向かうタスクであり，エージェントとランドマークはランダムな位置に複数生成される．各エージェントは他エージェントとの衝突を避けながらゴールであるランドマークに向かうことを目標とする．報酬設計は対象エージェントと全ランドマーク間における最小距離の負の報酬，衝突に対する負の報酬 (-1) とする．

4.2. 実験概要

Simple_spread を用いて MAAC と提案手法の学習を行い，最大報酬値と協調行動に対する定量的評価により比較する．学習条件は，エージェント数 3 (赤，緑，青)，ランドマーク数 3 (黒)，1 エピソード間の step 数を 25 とし，100,000 エピソード学習する．本実験では，他エージェントにランドマークを譲る行動を協調行動と定義する．この協調行動に対する定量的評価は，ランドマークが一定距離以上離れた環境において 100 エピソード間で対象エージェントが最も近いランドマーク以外へ移動した合計回数とする．この時，他エージェントは最も近いランドマークに移動するルールベース制御とする．また，Actor attention を可視化し協調行動に対する判断根拠を解析する．

4.3. 最大報酬と協調行動に対する評価

MAAC と提案手法の最大報酬値を表 1 に示す．表 1 から提案手法は MAAC と比較して最大報酬値が約 -1.5 低下している．協調行動に対する評価結果を表 2 に示す．表 2 から提案手法は MAAC より協調行動の合計回数が 56 回向上している．このことから Actor にアテンションヘッドを導入することで，最大報酬は低下するが，より良い協調行動を獲得できると考えられる．

表 1：最大報酬値

手法	最大報酬値
MAAC	-35.620
提案手法	-37.025

表 2：協調行動に対する評価

手法	赤	緑	青	合計
MAAC	50	63	69	182
提案手法	72	85	81	238

4.4. Actor attention の可視化

対象エージェントを赤と青とした場合の Actor attention の可視化例を図 2 に示す．可視化例は赤と青が協調行動を行ったシーンである．Actor attention の度合いを矢印の太さで表す．線が太いほど，Actor attention が強く注視していることを意味する．図 2 から step1，2，3 で赤と青は互いを注視している．ここから，赤は同じランドマークに向かう青を注視したことでランドマークを譲る行動をとり，青は赤を注視したことで赤の協調行動に合わせることができ，互いがランドマークに向かうことができたと考えられる．このように Actor attention を可視化することでエージェントの協調行動に対する判断根拠の解析を実現した．

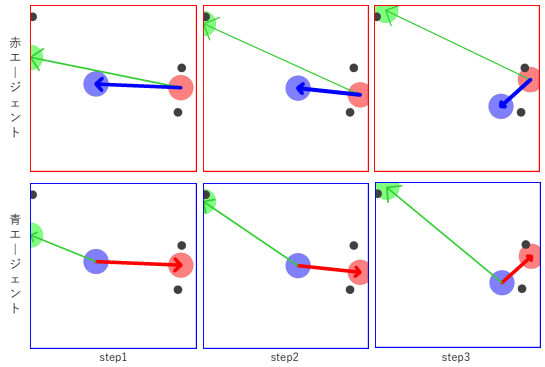

図 2：Actor attention の可視化

5. おわりに

本研究では，Actor-Critic 法の Actor に対しアテンションヘッドを導入した手法を提案した．評価実験から Actor にアテンションヘッドを導入することで，協調行動に対する判断根拠の可視化が可能となった．今後は，より良い協調行動の獲得と提案手法の汎用性の調査を行う．

参考文献

[1] J. Kim, *et al.*, “A cooperative multi-agent system and its real time application to robot soccer”, ICRA, 1997.

[2] S. Iqbal, *et al.*, “Actor-Attention-Critic for Multi-Agent Reinforcement Learning”, ICML, 2019.