

DA3: マルチエージェント深層強化学習における 協調行動の解釈性確立と対ノイズ性能の検証

元川 善就^{1,a)} 菅原 俊治^{1,b)}

概要: 本稿では、マルチエージェント深層強化学習におけるノイズの影響度を学習し、自律的に抑制する distributed attentional actor architecture model for multi-agent system (DA3) を提案する。ノイズは多くの分野に存在し、長年の間その特性の解明や抑制手法が考案されてきた。特に、周囲の限られた情報をもとに他のエージェントとの協調・調整、さらに相互協力を前提とした互惠行動を獲得する必要があるマルチエージェントシステムにおいても、ノイズ抑制や排除は重要な研究テーマである。本提案である DA3 は attention mechanism を内包しているため、エージェントは観測情報を取捨選択しながら環境に適応することが可能である。実験結果から、観測情報内のノイズや協調行動に無関係な第三者的エージェントに惑わされることなく、他エージェントとの協調行動を獲得し、attention mechanism を持たないベースライン(ここでは DQN) よりも高い学習性能を達成することを示す。また、DA3 内部に存在する attentional weights を解析することで、これまでブラックボックスとされてきたエージェントの行動決定過程におけるノイズの影響度などを把握できることを示す。

キーワード: マルチエージェント深層強化学習, 分散型自律学習, 注意機構, 協調行動解釈性, ノイズ抑制

1. まえがき

ノイズ抑制とは、様々な領域において取り組むべき重要な研究テーマである。長年、多くのノイズ特性の研究が続けられており、その特性に合わせたノイズ抑制手法が考案されてきた。たとえば、ロボティクスにおける微振動、電気回路における熱雑音、物理学における不確実性など研究領域ごとにノイズの定義や性質は異なるが、実世界では無視できないという点では共通であり、これは(深層)強化学習でも同様である。Wang ら [23] によると、強化学習におけるノイズは複数のカテゴリーに分類できる。例えば、*inherent noise* (内部雑音) はエージェントが情報を観測する際に、光線や熱雑音などの外的要因によって発生するノイズを指す。同様に、限られた局所情報から協調性を創発・学習するマルチエージェントシステム (multi-agent systems, MAS) あるいはマルチエージェント深層強化学習 (multi-agent deep reinforcement learning, MADRL) においても、システム全体の信頼性・堅固性を向上させるためにノイズ抑制は非常に重要な課題となる。しかし、発生するノイズ特性は想定する応用環境によって多種多様である

ため、既存研究のほとんどはノイズのない理想的な環境を仮定して学習エージェントのモデル化や協調性の獲得を検証することが多かった。

シングルエージェント深層強化学習の研究ではノイズを想定しているものもあるが ([4, 12, 19–21, 25] など)、多くの場合 *surrogate loss function* を利用するため、既存の強化学習アルゴリズムへの統合は容易ではない。MADRL においてはノイズを想定した研究は少なく、存在しても基本的にはエージェント間で共有する集中型ネットワークモデルでの学習 [5, 7, 17] を前提としている。また実世界で発生するほとんどのノイズの性質や発生場所は未知であるため、エージェントは学習前からノイズ特性が既知と想定する先行研究では実用性に欠ける。また、エージェントの行動場所やその処理能力・目的などによりノイズの影響度も異なるため、エージェント毎にノイズの特性や影響を自律的に把握し、それに惑わされず分散的に学習できる仕組みが必要である。

そこで本研究では、各エージェントが学習に不要なノイズの存在を同定し、そのノイズ抑制を考慮したマルチエージェント深層強化学習手法である *distributed attentional actor architecture for a multi-agent system* (DA3) を提案する。DA3 は *MAT-DQN* [10] と呼ばれる、*attention mechanism* を活用してエージェントの行動解釈性の確立

¹ 早稲田大学 基幹理工学研究科 知識ソフトウェア研究室, 東京

^{a)} y.motokawa@isl.cs.waseda.ac.jp

^{b)} sugawara@waseda.jp

を目指した学習モデルの拡張であり、これをノイズの同定に適用拡大している。DA3の特徴として、*saliency vector* と呼ぶパラメータの集合をネットワーク内部に導入し、行動決定過程における観測情報を超次元空間で表現する。本手法の重要な利点として、DA3を内部に保有するエージェント（以下 DA3 エージェント）へ観測情報のどの領域がノイズであるか、或いはどのエージェントが協調関係であるかという情報を事前に与える必要がない点である。例えば、あるエージェントの故障、異なる目的を持つエージェントとの共存、エージェントが持つセンサーの不良や、不良ではなくても距離に応じてセンサのデータにノイズが含まれ信頼性成果などの不確実性が発生する。このようなノイズを含んだ観測情報は直接 DA3 エージェントに与えられ、DA3 エージェントは観測情報のどの部分が（非）信頼できるかを自ら学習する。この観測情報は *saliency vector* として表現され、その表現に基づいて DA3 エージェントは行動を決定し学習する。

本研究ではノイズ環境下における DA3 エージェントの学習性能を検証・評価するために、*object collection games* を使用し、もっとも単純な *vanilla deep Q-network* (以下、DQN) との学習性能を比較する。実験結果から、DA3 エージェントはノイズを同定するとともに、そのノイズの混在比率に応じて信頼度を決定し、ノイズに惑わされることを最小限に抑え、その影響を下げる事が分かった。そして、比較手法よりも高い学習性能を達成できることが分かった。実際に、DA3 ネットワーク内の *attention mechanism* が学習した観測情報に対する *attentional weights* を分析した結果、エージェントの行動決定に強く関連する観測情報にはより大きな重みを与え、無関係なエージェントの情報やノイズを多く含む情報にはより小さい重みを与えるように学習していることを確認した。この解析結果は、DA3 エージェントはノイズに混乱することなく行動を決定できたかを可視化することも可能としている。

2. 関連研究

強化学習の分野ではノイズを考慮した研究は少ない [9, 11, 15, 18]。たとえば Raeisy ら [15] と Sabzevari ら [18] は *active noise control* (ANC) 問題を強化学習の枠組みで検証しているが、この ANC とは電気信号中のノイズをアクチュエータの集合で緩和する技術であり [6]、単音や複数音の周期的な正弦波音響ノイズにのみ着目している。Natarajan ら [11] は不偏推定と重み付き損失関数を用いてノイズ環境下におけるエージェントの耐性を検証した。他にも、不偏な *surrogate loss function* を定義することで、破損データやラベルノイズでの学習に関する研究もいくつかある [4, 12, 19–21, 25]。しかし MADRL の分野、特に各エージェントに与えられた分散型のネットワークではノイズを考慮した研究は筆者らの知る限り存在しない。また、

我々の目標は損失関数の変更やノイズの事前知識を一切与えずにエージェントがノイズ環境下で学習可能にする手法を提案することであるため、上記の先行研究とも目的が大きく異なる。

ノイズの存在は必ずしも有害とは限らず、有益なこともあることを主張した先行研究もある [2, 13, 24]。たとえば、Fortunato ら [2] はガウスノイズをネットワークの重みに意図的に追加することで、エージェントの環境探索を促進する *NoisyNet* を提案している。さらに Han ら [3] はエージェントの方策関数に確率性を導入した *NoisyNet* に拡張することで、エージェントの学習安定を実現した。しかしこれらの手法はエージェントの観測が常に正確でノイズを含まないことを前提としており、内在的な *inherent noise* などの観測情報のノイズがある場合は検証されていない。また Wang ら [23] はノイズによって変動のある報酬設計を提案し、強化学習モデルの堅固性向上を確認したものの、マルチエージェント環境での適用は想定されていない。

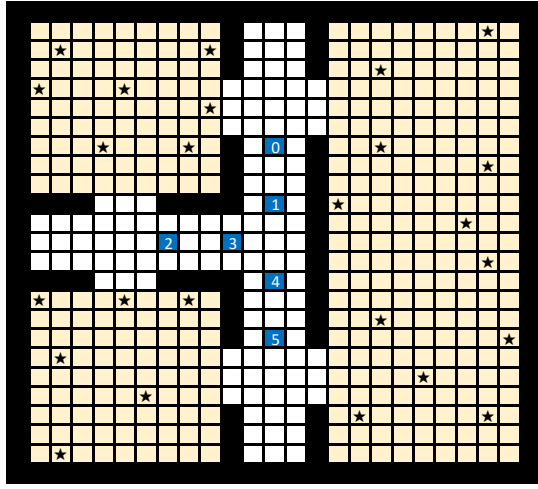
MADRL におけるノイズに関する研究は少ないものの、全エージェントで共有する集中型ネットワークを用いた学習モデルを想定したものが若干存在する [5, 7, 8, 16, 17]。しかし前述のとおり、エージェントの個体差や役割に応じて観測するノイズは異なり、またそれに対する影響や耐性もエージェントごとに異なる。一方、本稿で提案する DA3 は個々のエージェントが行動を学習する際に、観測情報のどの部分がノイズかノイズでないかを自律的に識別し、学習過程におけるノイズ発生による行動決定の混乱を低減することを検証する。また、エージェントが学習過程でノイズの多いデータや矛盾・不合理な行動をとる他のエージェントの位置を推定するために DA3 が役立つことを示すことで、DA3 エージェントの学習性能向上の説明を試みる。

3. 問題設定のモデル化

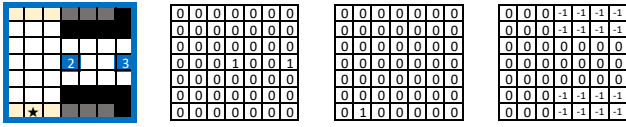
3.1 Decentralized POMDP

Markov decision process (MDP) とは、次ステップの状態が現在の状態のみに依存して遷移する動的システムの確率モデルのことを指す [14]。この時、 n 体のエージェントによる *decentralized partially observable Markov decision process* (dec-POMDP) を $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}, p_T, \{r_i\}, \{\Omega_i\}, \mathcal{O}, H \rangle$ で定義する。ここで要素は以下を表す。

- (1) $\mathcal{I} = \{1, \dots, n\}$: エージェントの有限集合
- (2) $\mathcal{S}$: 状態の有限集合
- (3) $\mathcal{A}_i$: エージェント i のとりうる行動の有限集合 ($i \in \mathcal{I}$)
- (4) $p_T(s'|s, a)$: 遷移関数 ($s, s' \in \mathcal{S}, a \in \mathcal{A}$)
- (5) $r_i(s, a)$: エージェント $i \in \mathcal{I}$ の報酬関数 ($s \in \mathcal{S}, a \in \mathcal{A}$)
- (6) Ω_i : エージェント $i \in \mathcal{I}$ の観測の有限集合 ($i \in \mathcal{I}$)
- (7) $\mathcal{O}(o|s, a)$: 条件付き観測確率関数 ($o \in \Omega, s \in \mathcal{S}, a \in \mathcal{A}$)
- (8) H : 確率過程の最大範囲 (≥ 0)



(a) 3 部屋の環境



(b) $N_C = 3$, $R_x = R_y = 7$ での観測情報のエンコード化

図 1: グリッド状の環境

このとき、エージェント i は割引累積報酬 $R_i = \sum_{t=0}^H \gamma^t r_i(s, a)$ を最大化するような方策 π_i を学習する。ただし、 γ を割引率 ($0 \leq \gamma < 1$) とする。本実験は決定的で離散的なパトロールタスクを想定しているため、離散的な時間ステップを t (≥ 0) と定義し、全ての確率関数は 0 か 1 のどちらかを取り得るとする。

3.2 問題設定と観測情報

本研究では図 1 に示すような $G_x \times G_y$ ($G_x, G_y \in \mathbb{Z}^+$, ここで $\mathbb{Z}^+$ は正の整数の集合) のグリッド状の環境を想定し、数体のエージェント (図 1a では数字の振られた青いマス) が移動しながら、環境で発生する回収対象物 (星型) を回収する object collection game を考える。各エージェントは、自身を中心とした $R_x \times R_y$ ($R_x, R_y \in \mathbb{Z}^+$) の範囲内を観測できる。各時間ステップにおいてエージェント $i \in \mathcal{I}$ は観測情報 $o_i \in \Omega_i$ を取得し、これを $N_C \in \mathbb{Z}^+$ チャネルの観測行列 (R_x, R_y) で表現し、内部のネットワークへ与える。観測行列は視野内の他エージェントの相対的な位置、回収対象物 (タスク)、壁や障害物などの情報を含む。エージェント $i \in \mathcal{I}$ は得られた観測情報をもとに行動 $a_i \in \mathcal{A}_i = \{\text{up, down, right, left}\}$ を決定し、実行する。なお $\mathcal{A}_i$ の各行動は隣接する上・下・右・左のいずれかへの移動を表す。移動先の事象によって以下の報酬を得るものとする。

- (1) $r_e > 0$: 対象物を回収 (同じ座標上に移動) して得られる報酬
- (2) $r_c < 0$: 壁やその他エージェントに衝突した場合に与

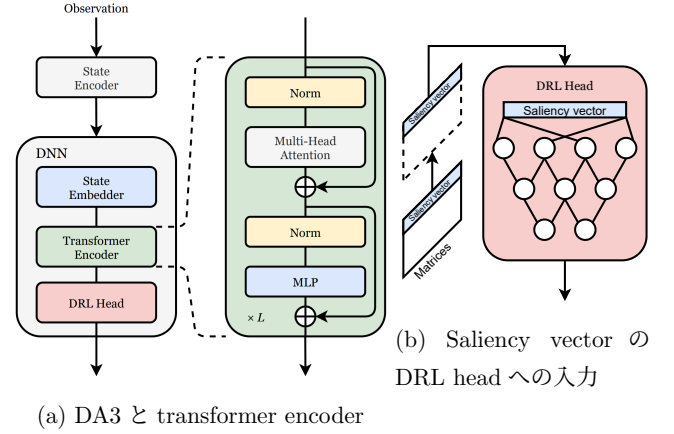


図 2: DA3 の構造

えられる負の報酬

- (3) $r = 0$: その他

各エージェント i は割引累積報酬 R_i を最大化するために衝突回数を減らし対象物を多く回収するような方策 π_i を自律的に学習する。なおこのゲームは単純な問題だが、本提案である DA3 の効果を確認するために他の要因を極力減らすために選択した。

3.3 自己注意機構 (Self-Attention)

Self-attention mechanism は attention mechanism の一種であり、観測行列内の異なるベクトル間の依存関係・類似性を示す [22]。Self-attention mechanism では、入力される系列から *query*, *key*, *value* の 3 つのベクトルが計算される。ここで *query* と *key* の各要素毎の類似性は、それぞれの内積によって求められる。その後 softmax 関数によって類似性を $[0, 1]$ 間に正規化して表し、これを attentional weights と呼ぶ。最後に *value* を掛けることで、以下の attention mechanism の計算が完了する。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V, \quad (1)$$

ただし、 Q, K, V , および d_k はそれぞれ *query*, *key*, *value*, そして *value* の次元数を示している。更なる詳細は [22] を参照していただきたい。

さらに、attention mechanism を h 個の attentional heads へ拡張し、*multi-head attention mechanism* を導入する：

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \\ \text{head}_l &= \text{Attention}(Q \cdot W_l^Q, K \cdot W_l^K, V \cdot W_l^V), \end{aligned} \quad (2)$$

ただし、 $W_l^Q, W_l^K, W_l^V, W_l^O$ は head_l ($1 \leq l \leq h$) とする際の学習対象パラメータ行列である。各 attentional head が異なる入力状態を特徴づけるため、 h の値を大きくすることで複雑な環境を効率よく学習することが可能になる。

4. 提案手法

提案手法である DA3 は全てのエージェントが独立したネットワークを所有することを想定している．各エージェントがもつネットワークの構造を図 2 に示す．この図が示すように，DA3 ネットワークには multi-head attention mechanism を含む *transformer encoder* が内在している．まずエージェントの観測データは *state encoder* で N_C チャネルの観測行列 (R_x, R_y) に変換（エンコード化）される．このとき各チャネルでは視野内にいるエージェント，回収対象物（タスク），また壁や障害物などの情報が 2 値 ($\{0, 1\}$ あるいは $\{0, -1\}$) で表現される．図 1b は図 1a におけるエージェント 2 の観測情報（図 1b の最左）と $N_C = 3$ のエンコード化した情報を示す．ここで，壁は黒色，壁向こうの観測できない領域は灰色で表現されている．左から 2 番目と 3 番目の行列は，エージェント 3 および対象物の相対的な位置が 1 で示され，4 番目の行列ではエージェント 2 の視野が壁などで観測不能な領域を -1 で埋めている．

カーネルサイズ $P \in \mathbb{Z}^+$ の畳み込みニューラルネットワークにより，エンコード化された N_C チャネルの観測行列は *state embedder* によって C チャネルの $\lfloor \frac{R_x}{P} \rfloor \times \lfloor \frac{R_y}{P} \rfloor$ 行列に展開される． P を *patched size* と呼ぶ．その後，展開された行列はさらに $\lfloor \frac{R_x}{P} \rfloor \cdot \lfloor \frac{R_y}{P} \rfloor + 1 \times C$ に平坦化される．このとき，長さ C の saliency vector を追加するため，1 が連結されている．Saliency vector は BERT [22] における *class token* に近い役割を担っており，vision transformer (ViT) [1] における *classification token* と同様の手法で実装される．しかし *class token* や *classification token* は入力情報を要約した形でベクトル化される一方，saliency vector は更に入力情報の顕著・重要な情報（協調エージェント，回収対象物など）を表現している．その後，ViT と同様に *position embeddings* が行列に付与される．

State embedder で得られた $\lfloor \frac{R_x}{P} \rfloor \cdot \lfloor \frac{R_y}{P} \rfloor + 1 \times C$ 行列を図 2a に示す通り *transformer encoder* へ入力する．このとき， $\lfloor \frac{R_x}{P} \rfloor \cdot \lfloor \frac{R_y}{P} \rfloor + 1 \times \lfloor \frac{C}{h} \rfloor$ の query, key, value がそれぞれ h 個計算される．観測情報の各要素と saliency vector の類似性は query と key の内積で計算され，その他の計算は先行研究 MAT-DQN [10] と同様である．Transformer encoder での処理は $L \in \mathbb{Z}^+$ 回行われるが，この値は環境の学習難易度によって決定するハイパーパラメータである．

最後に図 2b のように，*transformer encoder* から出力された行列から saliency vector のみを抽出して *DRL head* へ入力する．*DRL head* は学習対象に応じて変更することが可能であり（たとえば Double DQN, Dueling DQN など），エージェントの環境適応力を促進する狙いがある．図 2b の通り saliency vector のみが *DRL head* へ伝播するため，saliency vector と観測情報の要素との類似性を解析することで，エージェントが行動決定過程において着目・無視し

表 1: 使用されたハイパーパラメータ

概要	パラメータ	値
エージェント数	n	6
対象物回収時の報酬	r_e	1
衝突時の報酬	r_c	-1
エピソード長	H	200
エポック数		10,000
対象物数		25

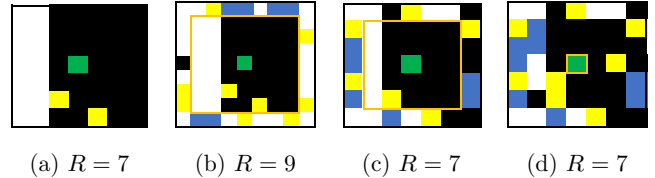


図 3: ノイズ発生時の同一観測情報の例. (a) ノイズなしの観測情報 ($R = 7$)，ただし中心の緑が観測するエージェント，黄が対象物，白が壁である．(b) Large marginal noise 発生時の観測情報 ($R = 9$)．(c) Small marginal noise 発生時の観測情報 ($R = 7$)．(d) Small full noise 発生時の観測情報 ($R = 7$)．いずれもオレンジ枠の外側でノイズが発生する．

ている情報を可視化することが可能になる．

DA3 エージェントは内部に *transformer encoder* を持つことで 2 つ利点がある．第 1 に，エージェントが観測情報の「取捨選択」が可能になる．つまり，エージェントが学習を経て有益な領域のみ着目（あるいは不要な部分を無視）するように観測情報を制限するため，効率的な環境学習に繋がる．第 2 に，エージェントの attentional weights を可視化することでこれまでブラックボックスだった行動決定過程に解釈性を持たせることが可能になる．前述の通り，saliency vector は観測情報を表現したベクトルであることから，エージェントの行動決定過程において影響のある情報が表現される．この点は先行研究 [10] でも認められたが，他にもノイズ環境下での attentional weights を解析することで学習に対するノイズの影響度なども多角的に確認できる．そこからの発見はシステムの更なる堅固性，効率化へと拡張でき，有意義と考えている．

5. 実験と結果

5.1 実験設定

DA3 エージェントのノイズ環境下での学習性能を検証するため，図 1a の環境にて object collection game を実行し，比較手法との差異を調査した．本実験の目的は，エージェントが環境観測時に inherent noise が発生した場合の学習性能の評価とノイズの影響を特定することである．この他に，ランダムに行動するエージェントを無視し，協調すべきエージェントを同定し，それらと協調行動を学習で

きるかを検証する実験も行ったが、ここではページ数の関係で省略する。実験に使用されたパラメータを表 1 に示す。以下に示す学習データは 20 試行の平均値である。

本実験では、各 DA3 エージェントは 4 つの attentional heads ($h = 4$) を所有し transformer encoder は 1 回のみ ($L = 1$) ループする。比較手法である DQN の構造は図 2 とほぼ等しいが、transformer encoder は含まれていないものとする。DA3 の DRL head には全結合層を採用し、一方、比較手法では 2 層の畳み込みネットワーク、max-pooling 層、そして全結合層とする。Patch size は $P = 1$ と設定し、観測情報の全ての要素が行動決定過程に関わるように学習させる。また、saliency vector の長さを $C = 64$ と設定した。エージェントの観測領域は一辺の長さが R ($= R_x = R_y$) の正方形と想定した。

実験は図 1a に示す 3 部屋のグリッド環境 (25×25) で実施し、ノイズ環境でのエージェントの振る舞いを検証した。環境内には 6 体のエージェントが存在し、それぞれが同じアルゴリズム (DA3 または DQN) で学習する。報酬設計として $r_c = -1$, $r_e = 1$ を設定した。観測情報は図 1b の通り $N_C = 3$ でエンコード化され、前述の通り各チャンネルでエージェント、対象物、壁あるいは観測できない領域が表現される。観測情報は直接エージェントのネットワークへと入力される。

各エピソードの開始時に、エージェントは環境中の固定番号位置に配置され (図 1), 25 個のオブジェクトがページ色の領域にランダムに発生する。エピソード長を $H = 200$ ステップとし、それまで、エージェントは環境を探索、学習する。エージェントが対象物を回収すると直後に別の対象物がページ色領域の重ならないランダムな位置に発生する。

5.2 観測情報におけるノイズ

本実験では、エージェントの観測情報にある確率で反転させるノイズを導入する。このノイズは [11] に基づくものであり、観測したエージェント、対象物、壁などが確率的に反転 (行列に応じて 0 と 1 あるいは 0 と -1) する。つまり、エージェントは観測している対象物の存在の有無に不確実性を持つことになる。図 3 にて観測情報におけるノイズの内容を示す。図 3a はあるエージェントのノイズのない観測情報 ($R = 7$) である。図の中央の緑は観測しているエージェント自身、青は観測領域内のその他のエージェント、黄は回収対象物、白は壁を表す。図 3b ($R = 9$) と図 3c ($R = 7$) は、センサーによる観測範囲のうち、観測可能限界付近の境界でノイズが高確率で発生する特徴を表した。これを *marginal noise* と呼ぶ (各図でオレンジ線の外側が確率に従って反転している)。また、観測範囲の境界のみならず領域内全体にノイズは存在するが、距離に応じて反転確率が徐々に小さくなる *full noise* を想定した場合

表 2: 各環境における学習性能の内訳

(a) DA3 による学習性能

発生ノイズ名	対象物回収量	エージェント間の衝突回数	壁との衝突回数
ノイズなし	271.45 $\pm$ 12.55	0.20 $\pm$ 0.12	2.98 $\pm$ 0.61
Large marginal	271.24 $\pm$ 11.88	0.24 $\pm$ 0.09	3.15 $\pm$ 1.11
Small marginal	208.59 $\pm$ 11.94	0.21 $\pm$ 0.10	2.47 $\pm$ 0.43
Small full	92.84 $\pm$ 6.96	2.56 $\pm$ 0.45	4.53 $\pm$ 0.52

(b) DQN (比較手法) による学習性能

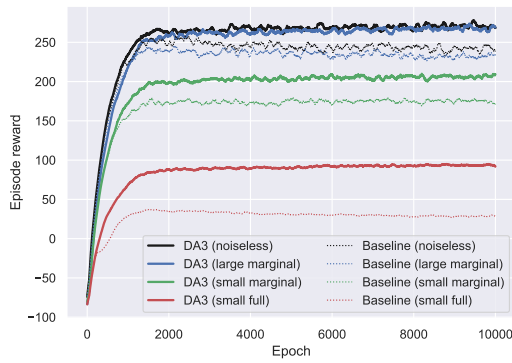
発生ノイズ名	対象物回収量	エージェント間の衝突回数	壁との衝突回数
ノイズなし	239.01 $\pm$ 14.78	0.32 $\pm$ 0.12	3.95 $\pm$ 0.70
Large marginal	233.31 $\pm$ 12.85	0.56 $\pm$ 0.17	4.46 $\pm$ 0.86
Small marginal	172.22 $\pm$ 11.20	0.29 $\pm$ 0.11	3.61 $\pm$ 0.52
Small full	29.41 $\pm$ 4.63	6.33 $\pm$ 1.39	13.49 $\pm$ 1.12

でも検証した (図 3d)。この full noise の場合、観測者自身に近い距離の情報は低い確率、遠い位置の情報は高い確率でノイズが発生するとする。なおデータの反転確率は、以下の実験結果の節で説明する。また、すべてのエージェントは同じ性能のセンサーを装備しており、ノイズの発生確率は同等と仮定した。

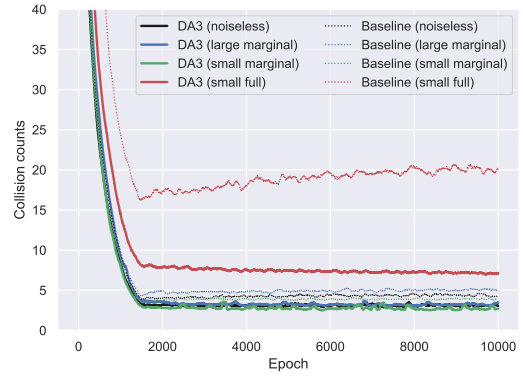
5.3 学習性能比較

DA3 エージェントが 3 種類の観測ノイズが発生する環境で (図 3b-d) に適応して学習できることを、ノイズなし環境での学習 (図 3a) と比較することで検証した。図 3b では、観測領域を通常より一回り大きく設定し ($R = 9$)、代わりに緑の情報が 0.5 の確率で反転するノイズ (以下、large marginal noise) を想定した。これは観測領域の緑に情報量は無いため、実質的には $R = 7$ のノイズなし観測情報と等価となる。しかし、エージェントはこのノイズに関する事実は与えられていない。図 3c では $R = 7$ とし、同様に緑部分の情報のみ 0.2 の確率で反転させる (以下、small marginal noise と呼ぶ)。つまり緑の情報は、信頼できる情報と出来ない情報がランダムに混在していることを意味する。最後に図 3d の small full noise 設定では、緑では 0.2、その内側では 0.1、さらにその内側 (観測エージェントの周囲) では 0.05 の確率で情報を反転させる。

図 4 にエージェントの学習推移に応じて、1 エピソードごとに獲得した報酬の推移と、エージェント同士・壁との衝突回数を示す。また、表 2 は実験の最後の 100 エピソード間における対象物の平均回収量、衝突の内訳を集計したものである。一般にノイズが増えるとその影響によってエージェントの学習性能はいずれの強化学習アルゴリズムに関わらず低下することが予想されるが、本実験でも同様の現象を確認した。ノイズなしの観測情報を得たエージェントが最も高い学習性能を達成し、特に small full noise のようなノイズが観測領域全体に発生する環境では学習性能が最も低いことが分かる。



(a) エピソード報酬



(b) 衝突回数

図 4: 観測情報にノイズが発生する際の学習性能比較

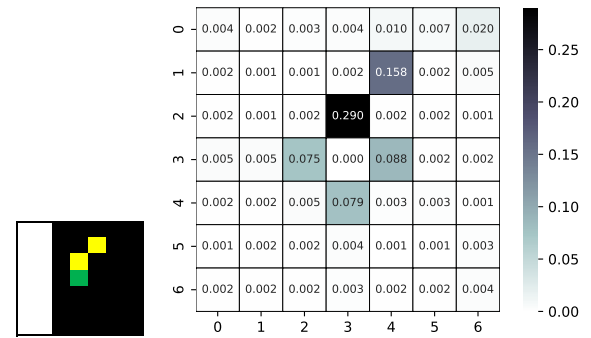
図 4 から, DA3 エージェントが比較手法より多くの対象物をより少ない衝突数で回収することができており, DA3 エージェントがどの環境下でも高い学習性能を達成したことが分かる. 表 2 を参照すると, 各環境におけるエージェントの対象物回収量の差はそれぞれ 32.44 (ノイズなし, 13.57% の性能向上) 37.93 (large marginal noise, 16.26% の性能向上), 36.37 (small marginal noise, 21.12% の性能向上), そして 63.44 (small full noise, 215.72% の性能向上) となっている. また, DA3 エージェントの衝突回数が比較手法よりも大きく抑制できたこと (特に small full noise) も確認できる.

さらに図 4 から, DA3 エージェントはノイズなしの場合と large marginal noise の場合でほぼ同程度の学習性能を達成している. このことは表 2a から確認でき, 対象物回収量 (差分は 0.21), そして衝突回数の差 (0.04) は僅かである. 実際に large marginal noise の場合は $R = 9$ であるが外周は事実上情報量がなく, ノイズ無しと同等の情報量であり, 提案手法ではこの事実を反映してほぼ同程度の学習性能を示した. 一方, 比較手法の DQN は, これらの間で性能劣化が確認された. その他のノイズの場合も DA3 エージェントが比較手法よりも高い学習性能を発揮しており, ノイズの頻度 (確率) に応じたノイズ抑制ができていえる.

5.4 Attention の解析・考察

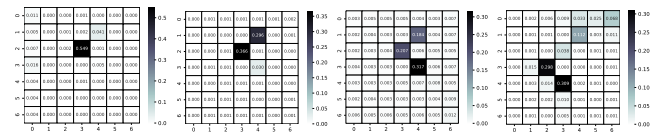
DA3 の transformer encoder 内部に存在する attentional weights から *attention heatmap* を作成することで, 観測情報のノイズが如何にエージェントの行動決定過程に影響を与えるかを解析した. 図 5 はノイズなし環境における (a) エージェントの観測情報, (b) 4 つの attentional heads ($h = 4$) からの平均 attention heatmap, (c) 各 attentional head の attention heatmap を表す.

ノイズなし環境での観測情報の例 (図 5a) からエージェントは左手に壁, 前方に 2 つの対象物を観測している. こ



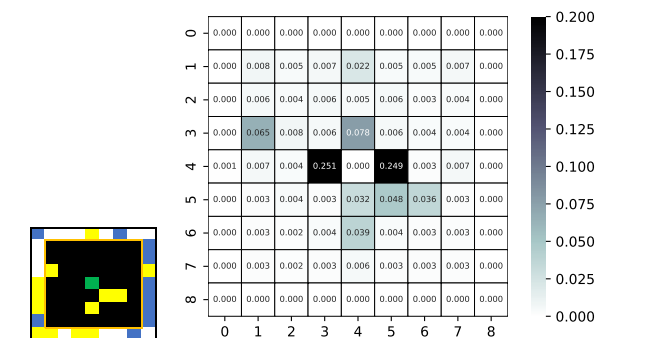
(a) 観測情報

(b) 平均 attention heatmap



(c) 各 attentional head ごとの attention heatmap

図 5: 観測情報と attention heatmap (ノイズなし)



(a) 観測情報

(b) 平均 attention heatmap

図 6: 観測情報と attention heatmap (large marginal noise)

のとき attention heatmap (図 5b) と照合すると, エージェントは自身の次位置の可能性のある四方と対象物に対して比較的強く着目していることが分かる. このエージェントは自身の上方に最も高い attentional weight を与えており, 上へ移動して即時報酬を得ることを期待していると考えられる. また, 図 5c より, 各 attentional head が異なる観測

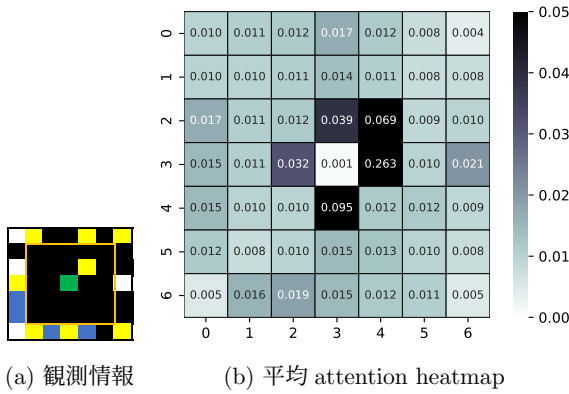


図 7: 観測情報と attention heatmap (small marginal noise)

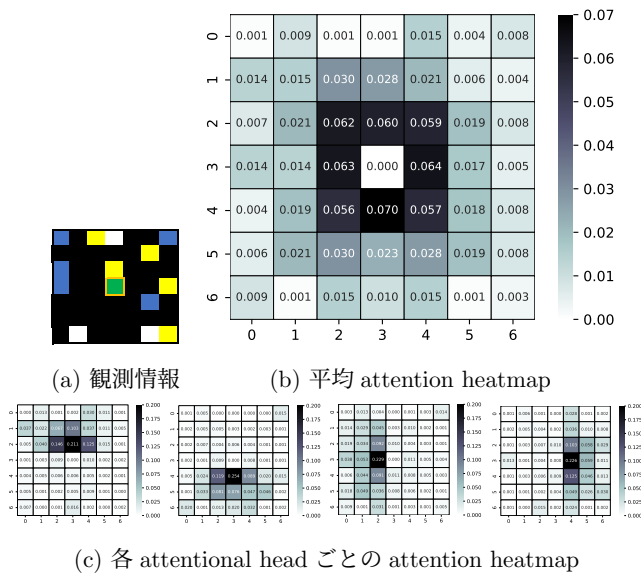


図 8: 観測情報と attention heatmap (small full noise)

領域に着目している。したがって h をある程度増やし、より効率的に環境を捉えた学習が可能になると思われる。この現象は MAT-DQN [10] でも確認されている。

次に図 6 は large marginal noise 環境での観測情報と attention heatmap である。先と同様にエージェントは自身の四方、そして対象物に対して着目していることが図 6a と図 6b から確認できる。一方ノイズの発生する観測領域の縁では、attentional weights がほぼ 0.000 となっていることも併せて確認できる。これは非常に顕著であり、エージェントが観測領域縁における情報量の皆無性を認め、意図的にその領域を無視するように学習できたと言える。DA3 エージェントは large marginal noise が発生する環境でもノイズなし環境での学習性能と遜色ない結果を達成しているが、この「ノイズを無視する」現象が要因と考えられる。

Small marginal noise 環境での結果である図 7 から、DA3 エージェントは、同様に自身の四方 (0.263, 0.039, 0.032, 0.095) と近くの対象物に比較的高く着目するように学習している。また、観測領域の縁における attentional weights の値は [0.005, 0.021] で推移しており、先の環境 (図 5b,

図 6b) での attentional weights の値と比較すると若干高い。これは small marginal noise に影響を受けている縁の情報を全く無視するわけではなく、むしろある程度考慮して次の行動を決定していると思われる。

最後に、small full noise 環境でのエージェントの観測情報と attention ヒートマップを図 8 に示す。図 8b からエージェントは自身の周辺 8 つを重点的に着目して行動決定することが示唆される。観測領域の縁に対する attentional weights はおよそ 0.007 程度であり、一回り中心に近い領域は 0.021、エージェントの周辺には 0.062 程度である。これはエージェントは遠くの対象物にはあまり着目せず、代わりに比較的信頼のできる観測領域 (自身の周辺 8 マス) に着目し、ノイズの発生率が高くなるほど低い attentional weights を与えるように学習しているからである。この特徴は先の環境で見られたエージェントの隣接する位置に対して強く着目する状況とは大きく異なる。これは近視眼的な行動となり、学習の性能をやや下げることになる。しかしその場合でも、DA3 エージェントはエピソード間に 100 以上の対象物を回収できた一方で、比較手法は 50 未満となり、その差は大きい。また興味深いことに、図 8c から各 attentional head がそれぞれエージェントの上下左右の一方を担当していることも確認できる。

5.5 考察

我々の実験結果から DA3 エージェントは、DA3 の部分で観測情報のどの領域が行動決定に重要かを判断しながら学習することで、ノイズの度合いに応じた処理をしていると思われる。実験では、DA3 エージェントは比較手法よりも優れた学習性能を示し、ノイズを含まないときも何を着目すべきかを学習し、観測情報にノイズが含まれる場合もその度合いに応じて重要さ (attentional weights) を学習し、ノイズに対する高い堅固性を示した。このような学習結果の説明は未知のことが多かったが、本研究によりエージェントの協調や競合回避行動の解釈性に役立つと考えられる。

6. むすび

本稿では、distributed attentional actor architecture model for multi-agent system (DA3) を提案した。提案手法によるノイズ抑制を検証するため、エージェントの観測情報にノイズを様々な確率で発生させた環境での学習性能をベースラインである (vanilla) DQN と比較した。結果、DA3 エージェントは行動決定に重要な情報に着目するように学習し、ノイズがある程度含まれる場合も、それに応じて重みを学習し、比較手法よりも優れた性能を達成できた。DA3 エージェントが保有する transformer encoder から attentional weights を可視化・解析したところ、観測情報に発生するノイズへの適応も分かるようになった。今

後の展開としてはノイズの特性の幅を広げ、現実に近い環境での検証へと拡張する予定である。

謝辞 本研究はJSPS 科研費 17KT0044 と 20H04245 の助成を受けたものです。

参考文献

- [1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houslsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *International Conference on Learning Representations* (2021).
- [2] Fortunato, M., Azar, M. G., Piot, B., Menick, J., Hessel, M., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C. and Legg, S.: Noisy Networks For Exploration, *International Conference on Learning Representations* (2018).
- [3] Han, S., Zhou, W., Liu, J. and Lü, S.: NROWAN-DQN: A Stable Noisy Network with Noise Reduction and On-line Weight Adjustment for Exploration, *CoRR* (2020).
- [4] Hendrycks, D., Mazeika, M., Wilson, D. and Gimpel, K.: Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise, *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R., eds.), Vol. 31, Curran Associates, Inc. (2018).
- [5] Kilinc, O. and Montana, G.: Multi-agent Deep Reinforcement Learning with Extremely Noisy Observations (2018).
- [6] Kuo, S. M., Kuo, K. and Gan, W. S.: Active noise control: Open problems and challenges, *The 2010 International Conference on Green Circuits and Systems*, pp. 164–169 (online), DOI: 10.1109/ICGCS.2010.5543076 (2010).
- [7] Li, S., Wu, Y., Cui, X., Dong, H., Fang, F. and Russell, S.: Robust Multi-Agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 4213–4220 (online), DOI: 10.1609/aaai.v33i01.33014213 (2019).
- [8] Lin, W., Zhixin, L. and Lei, G.: Robust Consensus of Multi-agent Systems with Noise, *2007 Chinese Control Conference*, pp. 737–741 (online), DOI: 10.1109/CHICC.2006.4347503 (2007).
- [9] Minutti, C., Gomez, S. and Ramos, G.: A machine-learning approach for noise reduction in parameter estimation inverse problems, applied to characterization of oil reservoirs, *Journal of Physics: Conference Series*, Vol. 1047, p. 012010 (online), DOI: 10.1088/1742-6596/1047/1/012010 (2018).
- [10] Motokawa, Y. and Sugawara, T.: MAT-DQN: Toward Interpretable Multi-agent Deep Reinforcement Learning for Coordinated Activities, *Artificial Neural Networks and Machine Learning – ICANN 2021* (Farkaš, I., Masulli, P., Otte, S. and Wermter, S., eds.), Cham, Springer International Publishing, pp. 556–567 (2021).
- [11] Natarajan, N., Dhillon, I. S., Ravikumar, P. K. and Tewari, A.: Learning with Noisy Labels, *Advances in Neural Information Processing Systems* (Burgess, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. and Weinberger, K. Q., eds.), Vol. 26, Curran Associates, Inc. (2013).
- [12] Patrini, G., Rozza, A., Menon, A., Nock, R. and Qu, L.: Making deep neural networks robust to label noise: a loss correction approach, *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 2233–2241 (online), DOI: 10.1109/CVPR.2017.240 (2017).
- [13] Plappert, M., Houthoofd, R., Dhariwal, P., Sidor, S., Chen, R. Y., Chen, X., Asfour, T., Abbeel, P. and Andrychowicz, M.: Parameter Space Noise for Exploration, *International Conference on Learning Representations* (2018).
- [14] Puterman, M. L.: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., USA, 1st edition (1994).
- [15] Raeisy, B. and Golbahar Haghighi, S.: Active Noise Controller with reinforcement learning, *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, pp. 074–079 (online), DOI: 10.1109/AISP.2012.6313721 (2012).
- [16] Raeisy, B., Golbahar Haghighi, S. and Safavi, A. A.: Active Noise Control System via Multi-Agent Credit Assignment, *J. Intell. Fuzzy Syst.*, Vol. 26, No. 2, p. 1051–1063 (2014).
- [17] Ryu, H., Shin, H. and Park, J.: Multi-Agent Actor-Critic with Hierarchical Graph Attention Network, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 7236–7243 (online), DOI: 10.1609/aaai.v34i05.6214 (2020).
- [18] Sabzevari, S. A. H. and Moavenian, M.: Application of reinforcement learning for active noise control, *Turkish J. Electr. Eng. Comput. Sci.*, Vol. 25, pp. 2606–2613 (2017).
- [19] Scott, C.: A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (Lebanon, G. and Vishwanathan, S. V. N., eds.), Proceedings of Machine Learning Research, Vol. 38, San Diego, California, USA, PMLR, pp. 838–846 (2015).
- [20] Scott, C., Blanchard, G. and Handy, G.: Classification with Asymmetric Label Noise: Consistency and Maximal Denoising, *Proceedings of the 26th Annual Conference on Learning Theory* (Shalev-Shwartz, S. and Steinwart, I., eds.), Proceedings of Machine Learning Research, Vol. 30, Princeton, NJ, USA, PMLR, pp. 489–511 (2013).
- [21] van Rooyen, B. and Williamson, R. C.: Learning in the Presence of Corruption (2015).
- [22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Vol. 30, Curran Associates, Inc., (2017).
- [23] Wang, J., Liu, Y. P. and Li, B.: Reinforcement Learning with Perturbed Rewards, *AAAI* (2020).
- [24] Xu, H., Zhang, C., Wang, J., Ouyang, D., Zheng, Y. and Shao, J.: Exploring Parameter Space with Structured Noise for Meta-Reinforcement Learning, *IJCAI* (2020).
- [25] Zhang, Z. and Sabuncu, M.: Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels, *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R., eds.), Vol. 31, Curran Associates, Inc. (2018).