人工知能学会研究会資料
SIG-AGI-008-03

エピソード記憶と価値を紐づけた海馬モデルによる行動学習の分析

Analysing Behavior Leaning of Hippocampus Model that Connects Episodic Memory and Value

堤 優奈[1]　栢沼 晋太郎[1]　川添 紗奈[1]　宮田 真宏[2]　大森 隆司[1]

Yuna Tsutsumi[1], Shintaro Kayanuma[1], Sana Kawazoe[1], Masahiro Miyata[2], and Takashi Omori[1]

[1] 玉川大学 工学部
[1] College of Engineering, Tamagawa University.
[2] 玉川大学大学院 工学研究科
[2] Graduate School of Engineering, Tamagawa University.

Abstract: Episodic memory is the important function of hippocampus, and we can't lack it for our wholesome life. But its theory and role in decision making is not clear yet. In this study, we analyze and discuss on its model that we proposed at the WBAI hackathon. As the result, an association of the episodes and the value enabled rapid action learning with small number of experiences, and takes complementary role with the strong but slow feature of Deep Learning.

1. はじめに

エピソード記憶は，人間の重要な認知機能の一つであり，その人物の生活史とアイデンティティを支える記憶体系である．古来，記憶現象についての研究は多くあるが，脳科学ではその機能の責任部位として海馬が知られており，その損傷は健忘症を引き起こす[1][2]．人間の知能という側面からみると，エピソードはアイデンティティだけでなく行動決定による生存率向上にも間違いなく重要であろうと思われるが，実際場面におけるその役割は現時点では明確ではない．

エピソード記憶についての理解がそういう状態であるときに，2017 年 9 月，NPO 法人全脳アーキテクチャ・イニシアティブ主催で，第 3 回全脳アーキテクチャ・ハッカソン「目覚めよ海馬！ 汎用人工知能プロトタイプに向けた海馬モデルの組み込み」が開催され[3]，全 8 チームがマウスの行動実験における迷路のシミュレーション課題に取り組んだ．我々はこれに参加し，エピソード記憶に価値を連合させてオンラインで意思決定させるアルゴリズムを提案・実装した結果，準優勝であった．ハッカソンでは，1 次元の仮想迷路として 8 個のタスクが用意され，そこで行動するエージェントのプロトタイプが与えられて，それを改造してタスクを解くことが求められた．課題の公表からハッカソンまでの二カ月の間の試行錯誤の結果として，我々の海馬モデルは 1 次元仮想迷路課題 8 タスクのうち 7 タスクを単一エージェントで連続して解決する成績が出せた．しかしその内部で何が起きていたのか，エピソードに価値を連合されたことは有効だったのか，それはなぜか，など不明な点が多く残るままにハッカソンは終わった．

そこで本研究は，我々が提案した行動決定アルゴリズムの内部過程を分析し，なぜこれがタスク解決，特に徐々に難しくなるタスク群のすばやい解決に有効であったのかを明らかにすることを試みる．そして，従来から知られている行動決定アルゴリズムとの関係を議論することで，機械学習におけるエピソード記憶や価値の連合の効果について議論する．

2. エピソードと価値の紐づけによる行動決定過程のモデル化

2.1. ハッカソン

課題は，全長 24m，幅 2m の仮想迷路でエージェントが各タスクで事前に決められた条件をクリアすると報酬が得られ，解決できたトライアル数が一定条件を満たすと次のタスクへ進むものであった（Fig.1）．基本的にはエージェントが緑色のブロック

を取るとそのトライアルは成功となるが，同時にエージェントは常に負の報酬を受け続け，累積報酬がある値を下回った時点でそのトライアルが失敗となる．あるタスクで，成功数が失敗数よりも 23 回多くなったなら，それまでの学習結果を維持したまま次のタスクに進むことができる．1 次元迷路のタスク 1～7 の特徴を表 1 に挙げる．

Fig.1　1 次元の仮想迷路，エージェント(水色の玉)がゴール(緑色のブロック)に到達すると報酬が与えられ，トライアルが終了する．迷路の途中に壁に色のついた部分があり，タスクごとにその場所でのタスクの要請が変化する。）

表 1. 1 次元仮想迷路でのタスク一覧

タスク 1	無条件で，ゴール地点で報酬が与えられる．
タスク 2	緑の壁の通過時に(すぐにその場で) 報酬が与えられる．その後，ゴールでも報酬が与えられる．
タスク 3	緑の壁の地点で 2 秒間待機すると，報酬が得られ，さらにゴールでも報酬が得られる．緑の壁の地点で待機せずにゴールに到達しても報酬は得られない．
タスク 4	報酬が得られる壁が赤の壁の地点に変わる．報酬が与えられる条件はタスク 3 と同じ．
タスク 5	報酬が得られる壁が青の壁の地点に変わる．報酬が与えられる条件はタスク 3 と同じ．
タスク 6	「タスク 4→タスク 5→タスク 3→タスク 4・・・」と報酬が与えられる地点が毎回変化する．
タスク 7	スタート時に幾何学模様(△，○，□)をエージェントに提示される．△ならタスク 3 と同じように報酬が与えられる．○ならばタスク 4 と同じ，□ならタスク 5 と同じように報酬が与えられる．

エージェントには RGB カメラと深度カメラが搭載されており，行動ステップごとに画像を取得する．ハッカソン参加者はエージェントのプログラムを作り，この画像から何らかの処理をして行動を決定する．エージェントの行動は前進(1m/step)，右回転(+10 度回転/step)，左回転(-10 度回転/step)，停止(±0/step)の 4 種類であり，結果としてエージェントはその 2 次元的位置と回転方向の状態量(x, y, θ)を持つ[4].

2.2 エピソード記憶と価値による意思決定

我々は，人の海馬のエピソード記憶の機能とげっ歯類の海馬の場所細胞の機能に注目し，行動学習モデルを提案した．エピソード記憶は，視覚・聴覚・触覚などの感覚情報の処理の結果として海馬に与えられる特徴の集合体である特徴ベクトルとした．さらに本研究では，エピソードには多くの場合に情動が付随すると考える．例えば，道を歩いていてかわいい犬に会って嬉しかった場合は快の情動(正の価値)がエピソードに付随し，犬に吠えられて恐かった場合は不快な情動(負の価値)がエピソードに含まれる．これより，我々はエピソード記憶には価値が紐づけられると考え，我々の提案モデルではエピソード記憶には現在の感覚情報の特徴ベクトルの一部に価値情報を付随させ，それがエピソードに基づく簡易な意思決定を可能にしている．例えば，分かれ道で左に行けばかわいい犬がいて，右に行けば吠える犬がいるという記憶があれば，私たちは左に行くことが多いだろう．これは，過去のエピソードを想起したときにその価値が同時に想起され，過去の経験を現状に当てはめた結果としての未来の価値が予測されて行動選択がなされた[5]，と説明できる．

さらに，このタスクはナビゲーションタスクという性格上，場所ごとに適切な行動が存在する．そのため，エージェントが自己位置を特定できるならば，その場所での行動の価値を過去のエピソードから高精度で想起・予測できると期待する．そこで，海馬にある場所細胞の機能として視覚情報の時系列中で変化の少ない部分を抽出して自己位置を特定する機能をもつ Slow Feature Analysis(SFA)を実装した[6].
以上より，本稿では上記のエピソード記憶のみで行動決定するモデル A と，エピソード記憶と場所細胞の組み合わせで行動決定するモデル B の 2 つのモデルを検討した．

Fig.2 は本モデルの基本的な構造である．エピソードは，視覚入力 V，後述の SFA 出力 S，その時の行動 A，事後に付与される行動の価値 R からなる．モデル A は認識系から視覚情報の特徴ベクトル X を受け取り，エピソード記憶として溜められている過去の特徴ベクトル群 E^t より，特徴ベクトルの近い近傍エピソード群 Z を選ぶ[1].

モデル A の計算過程は以下のようになる．

[1] 特徴ベクトル間の距離に何を使うかは問題に依存し，本稿ではその点の分析には及んでいない．

エピソード ： $E^t = \{V_e^t, A_e^t, R_e^t\}$
行動の選択肢 ： $A^t = \{a^i : i = 1 \ldots 4\}$
感覚入力： $X = \{V_x\}$
近傍エピソード群 ： $Z = \{E^t : \|V_x - V_e^t\| < Thr\}$

モデル B には感覚入力として視覚入力に加えて SFA で検出された特徴量を用いた．SFA は、時系列中で急速に入力信号が変化する中で、変化の少ない特徴を抽出するための教師なし学習アルゴリズムである．例えば、道を歩いている時、遠くの山は自分から見てもあまり変化がない．しかし、目の前にあるモノは目まぐるしく変化を起こす．これより，変化の少ない感覚入力は自己の移動に安定した情報となり，おそらくは自己位置の推定の指標になると期待できる．モデル B は、この変化の少ない視覚特徴を抽出することで場所を判断しようとした．具体的には，モデル B では視覚入力から元の視覚情報の特徴ベクトルに加えて長さ 32 の場所を示す特徴量 S を作り出し、それをエピソード X として溜めた過去の特徴ベクトル群 Et より類似したエピソード群 Z を選んだ．それ以降の計算過程はモデル A と同様な処理を行ない、各場所での行動選択を試みた．

エピソード ： $E^t = \{V_e^t, S_e^t, A_e^t, R_e^t\}$
行動の選択肢 ： $A^t = \{a^i : i = 1 \ldots 4\}$
感覚入力＋SFA 出力：$X = \{V_x, S_x\}$
近傍エピソード群 ：
$Z = \{E^t : \|[V_x, S_x] - [V_e^t, S_x]\| < Thr\}$

選ばれた近傍エピソード群を行動ごとに分類し，各行動の価値の平均を求めることで入力特徴ベクトルに対する行動価値を計算し，SoftMax 法で行動を選択する．同時に，現在の特徴ベクトルと環境から与えられた報酬をエピソードとして保存する．

Z の行動毎の分類 ： $Z^i = \{E^t : E^t \in Z, A_e^t = a^i\}$
行動 i の価値 ： $V^i = \sum_{E^t \in Z^i} R_e^t \Big/ |Z^i|$
行動の選択 ： $A_x = SoftMax_i V^i$
保存するエピソード ： $E^{new} = \{V_x, S_x, A_x, R_x\}$

行動した結果として最後の報酬を得た場合はそのトライアルが終わり，そのトライアル内のエピソードの価値情報を過去に遡って更新することで，最終結果に応じた価値 R^t をエピソードに付与する．エピソード記憶は過去 10,000 ステップを保持しているが，その保持数に関しては検討が必要である．

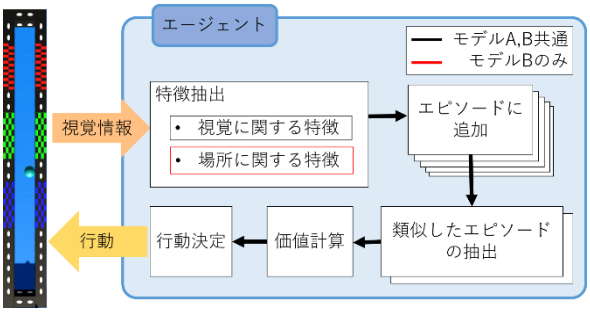

Fig.2 エピソード記憶による意思決定モデルの構造

2.3. 学習のパフォーマンスと場所細胞の導入の効果

本ハッカソンでサンプルとして提供された行動学習モデルでは，経験した状態，行動，報酬などをメモリに蓄積し，ランダムサンプリングする Experience Replay を用いて行動決定している[7]．この方法では経験の時系列を学習に利用できないためか，一定の場所で待機するという動作の学習が成功せず，タスク 3 で行き詰った．しかし，我々が提案した現在状態に類似したエピソード記憶の価値から行動選択するモデルＡは，一定の場所で待機することの学習に成功し，サンプルでは解けなかったタスク 3 を始めタスク 7 までをクリアできた．これには，エピソード記憶に価値を紐づけて未来の価値を予測する方式が効果を持った可能性がある．

一方で問題となったのが，エージェントが迷路内で不必要に回転してフィールド内を無駄に往復して起こる，タスクのクリアまでのステップ数の多さである．これに対して我々は，エージェントが迷路を不必要に往復するのは Fig.1 の黒い壁の方をエージェントが向いている時の視覚情報だけから類似エピソードを選ぶモデルＡでは，今いる場所が正確に識別できない，との仮説を立てた．そこで，場所を特定する場所細胞の機能を SFA で実現・実装したモデルＢを用意した．SFA の効果は大きく，タスク 1 から始めてタスク 7 をクリアするまでにかかったステップ数は，SFA を実装しないモデルＡの約 40,000 ステップに対し，SFA を実装したモデルＢは 23,000 ステップ強であった．Fig.3 より各タスクのクリアまでの効率が大きく向上していることがわかる．

では，場所細胞を実装しないモデルＡと実装したモデルＢの振る舞いはどう違うのか．次節ではそれ

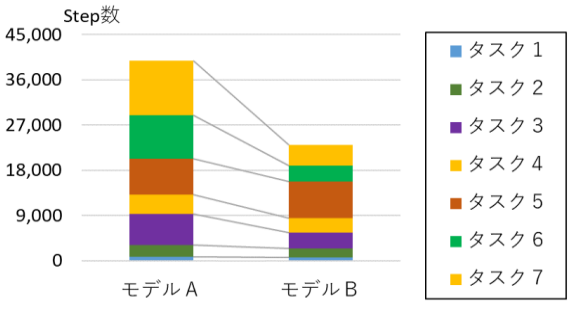

Fig. 3 場所細胞がなしのモデル A とありのモデル B でのタスク 7 終了までのステップ数の差

ぞれのモデルの内部でなにが起きていたか，エージェントの動きを分析する．

3. モデルの内部過程の分析

3.1 エージェントの軌跡からの分析

我々は，提案モデルの解析を行なう場面としてタスク 3 に着目した．タスク 1 とタスク 2 は直進をするだけで報酬が得られたのに対して，タスク 3 は直進以外の回転や停止などの行動により緑の壁の前で一定時間留まってはじめて報酬が得られる．つまり，直進だけでなく，場面に応じて回転・停止などの行動の選択が必要であった．この行動選択をどのように獲得してタスクをクリアしたかを明らかにすることで，我々の提案したエピソードと価値を紐づけた行動選択モデルの特性が明らかになると期待する．

まずモデルAがタスク 3 をクリアする過程を分析した．価値を含んだエピソードの検索は，過去の類似場面での価値から未来の価値の予測につながる．そのため，適切なエピソードの蓄積を通じたモデルの学習によりエージェントの行動がどう変化したか確認した．具体的には，個々の場所に対する行動確率分布を計算し，学習によるその変化の傾向を見た．

可視化対象のエピソードは，タスク 3 で序盤に報酬を得たグループと終盤に報酬を得たグループから，ステップ数が近いものを選んだ．そして，各ステップの行動選択を，その進行方向(z 軸)の位置に対応して Fig.4, Fig.5 に示した．また，タスク中で一定時間留まる必要がある壁の緑色部分を緑の背景で示した．

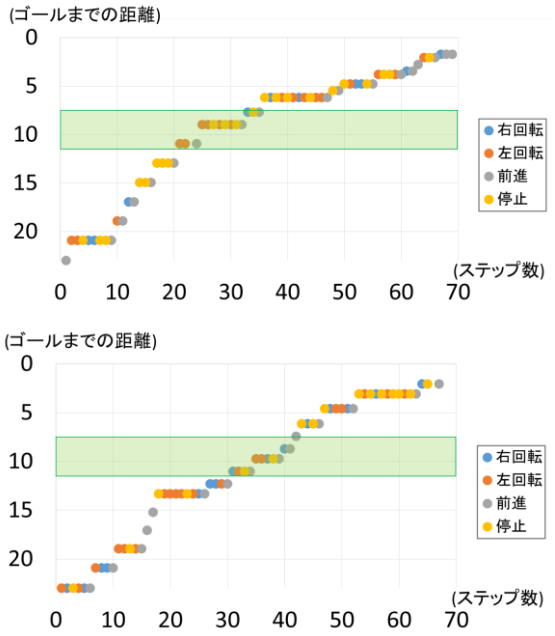

Fig. 4 タスク 3 の序盤(a：上段)と終盤(b：下段)から抽象したトライアルの進行状況

Fig.4(a：上段)では，緑の壁付近で左回転と停止の頻度が高く，壁の前で待てていた．しかし，特定の位置で長く停止していることから，偶然に直進行動が出なかった可能性が高い．それに対し Fig.4(b：下段)では可能な行動がバランスよく選択されて緑の壁の範囲で停止していることから，その場にとどまりやすい行動確率が学習できているようにみえる．

そこで，価値の高いエピソードの蓄積による行動確率の変化を分析した．そのために，タスク 3 全体をスタート位置から緑の壁の前まで，緑の壁の範囲，緑の壁の後から報酬が得られる位置までの 3 区間に分け，そのそれぞれについて行動確率の平均値を計算した(Fig.5(a)(b))．これより，当初は右回転と左回転の確率がアンバランスで左向きの回転が起きる傾向が強かったが，価値の高いエピソードの蓄積により行動全体で右回転と左回転の確率が均衡して進行方向がゴール方向からぶれる可能性が減り，さらに緑の壁の範囲では左右の回転の確率が高くなって前進の確率が下がり，結果としてこの区間の滞在時間が長くなるように変化していることがわかる．

以上より，モデルAはエピソードの蓄積により行動学習ができることが確認できた．また，場所に対応して実際に行った行動の割合だけでなく，行動確率も場所に準じて変動し，タスク 3 の間に効率の良い行動が選ばれやすくなっていることが確認された．

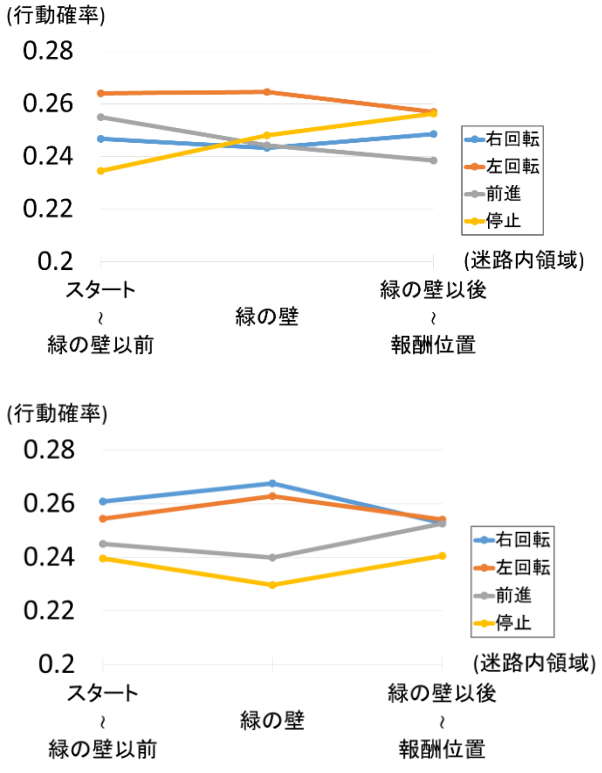

Fig5. 序盤トライアル(a：上段)と終盤トライアル(b：下段)の行動確率の平均値

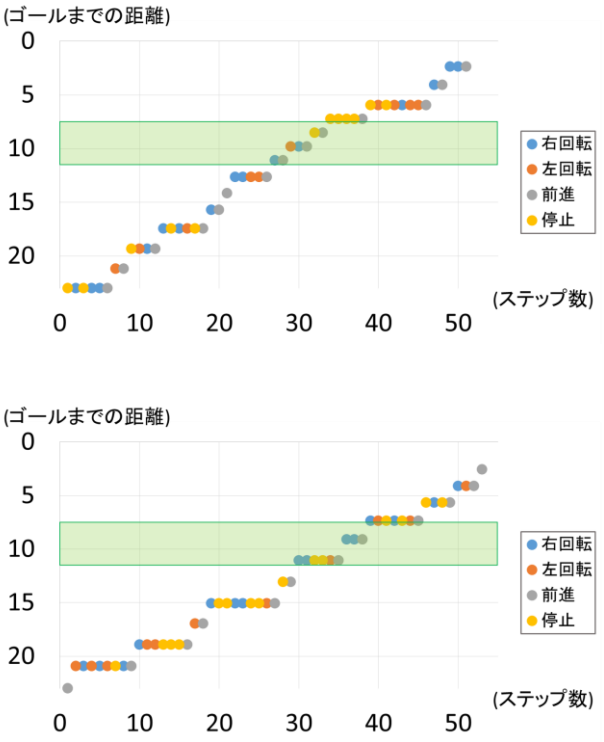

Fig. 6 序盤付近のトライアルの経路(a: 上段)と終盤付近のトライアルの経路(b: 下段)

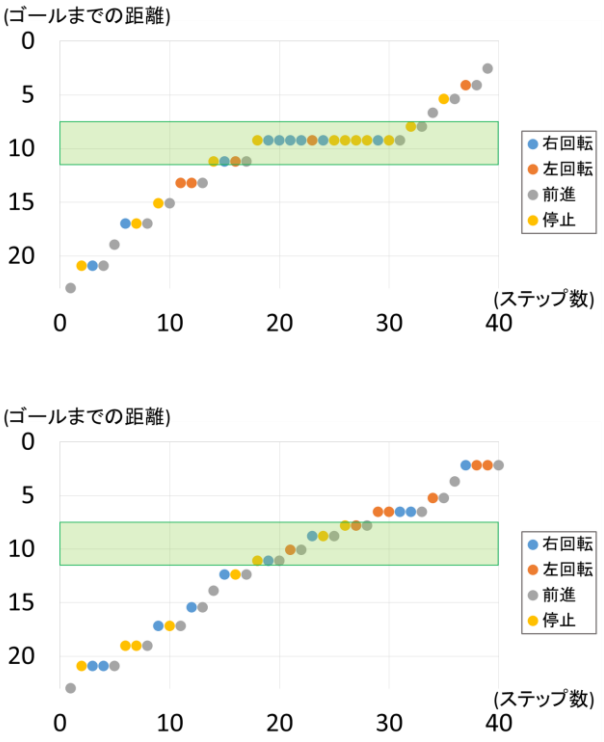

Fig. 7 タスク 2 序盤のトライアルの経路(a:上段)とタスク 2 終盤のトライアルの経路(b:下段)

3.2 なぜ場所細胞は効果があるのか

次に，モデルＡに SFA を実装したモデルＢで，効率が格段に向上したメカニズムを分析した．そのため，前節と同様の可視化を行った(Fig.6 (a),(b))．これをモデルＡの場合 (Fig.4)と比較すると，トライアルの序盤と終盤での行動の変化が少ないように見える．これより，タスク 2 で報酬を獲得したエピソードにおいてすでに緑の壁の範囲内で停止・回転の行動をとっている可能性を考え，同様の可視化をタスク 2 でも行った(Fig.7 (a),(b))．

Fig.7 (a),(b)から，モデルＢではタスク 2 の時点で，緑の壁の前でたまたま長く滞在して，それからさらに前進してゴールで報酬を得た，というトライアルがあったと推測される．そういう成功トライアルがあると，以降はその行動を再現するのが本モデルの特性である．すなわち，タスク 2 の段階でタスク 3 に通じる行動学習を行ったことが，タスク群を通じて効率があがった一因であったと推測できる．

では，なぜ SFA の導入が学習を加速したのだろうか？一つの仮説は，位置を特定する精度がモデルＢにて高まったということである．それを検証するため，現在の入力エピソードに近いとして抽出された過去のエピソードの位置関係を可視化した(Fig.8)．図中の黄色の星が現在のエピソードの位置で，その他の丸が抽出されたエピソードの位置である．上段では，現在の位置に対して遠い位置に類似エピソードが出ているが，下段では近い位置・角度のエピソードだけが抽出された．また，類似とされた個数が上段では 5 個，下段では 2 個と，蓄積されたエピソードの質の影響もありうるが，SFA ありの方が明らかに近い場所のエピソードを検出していた．

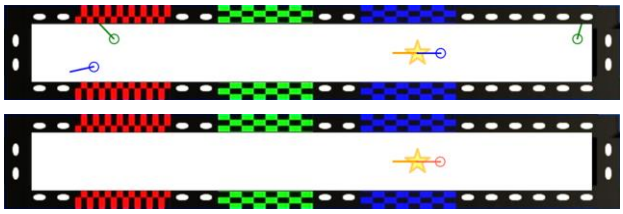

Fig. 8 SFA を実装しない場合(上段)と実装した場合(下段)の類似トライアルの位置

4. 考察

エピソード記憶に価値を紐づけることによってタスク 1 からタスク 7 まで一つのエージェントが連続してクリアできた．またエージェントの内部計算では，過去の経験から未来の価値がある行動を選択できた．しかし，ハッカソンの課題で与えられた感覚入力のままのエピソード情報では現在の位置を正確

に判別できず，適切なエピソードが選べず，結果として未来の価値予測が狙い通りには働かなかったと思われる．そしてこの問題は，SFA による場所細胞を用いることで大きく改善された．これは，エージェントが現在の位置に関係するエピソードを記憶から的確に選ぶことで現状にマッチした行動ー価値関係の計算が可能となり，未来の価値が高いエピソードを精度よく現在に当てはめることができるようになったことによると考えられる．

本ハッカソンのナビゲーションという課題は，場所に依存して行動価値が決まるものであり，エピソード記憶の「どこ」という情報の精度よい認識が成功の鍵となった．同ハッカソンで 1 位となったチームはその学習に強化学習を用いており，その中の特徴抽出部が本研究の SFA と同じ機能を実現したと考えられる．一方でタスク 1 からタスク 7 までを連続して過去の学習結果を再利用しながら高速に解いていくという効率性からは，場所細胞とエピソード記憶を用いる本モデルは効率的であると考えられる．実際, タスク 3 については 48 トライアルで完了しており，より困難な課題があるタスク 1 からタスク 7 までの全体でも 361 トライアルで完了している．これは，極端な場合には 1 回の成功トライアルで次からは適切な行動がとりうる本モデルのメリットである．一方，初期に悪いパターンで失敗が続くという可能性は排除できず，その場合には後の経験で回復できるとは思われるが，その評価までは至っていない．

5. まとめ

以上，エピソード記憶に価値を紐づけした行動決定モデルのメカニズムについて，全脳アーキテクチャ・イニシアティブのハッカソンの課題を例に分析した．この課題の特性は，タスク 1 からタスク 7 まで順次に，それまでの学習結果を活かしたまま次の課題に取り組むという，知識の蓄積に相当する機能を要求するものであった．このような課題は現実世界では珍しくはないが，DQN を代表とする Deep Learning による行動学習では学習に時間がかかることが予想され，ここに Deep Learning の一つの限界があるよう思える．それに対してエピソード記憶を用いる方法は，状況変化や新規の経験にすばやい解決を与える意味で，効果的であろう．

大量のデータに基づき特徴抽出から行動決定までを統合的に学習する Deep Learning は，時間をかけてよいのであれば強力である．逆にエピソードによる方法は事例の揺らぎに強く依存し，最適性を求めるのは難しい．両者を統合して，データや経験が増えるにしたがってエピソードベースから階層ネットワークベースに移行していく方式の開発が求められよう．これについては，今後の課題としたい．

謝辞

本研究は全脳アーキテクチャ・イニシアティブのハッカソンに大きく触発された．このような問題に取り組む機会をいただけたことに感謝する．また，SFA の実装は早川博章先生の支援による．深く感謝する．

参考文献

[1] スザンヌ・コーキン：ぼくは物覚えが悪い 健忘症患者 H.M の生涯，早川書房,2014
[2] スクワイヤ，カンデル：記憶のしくみ 上・下，講談社，2013
[3] 第 3 回全脳アーキテクチャ・ハッカソン，https://wba-initiative.org/2391/
[4] 第 3 回全脳アーキテクチャ・ハッカソン 迷路課題仕様 ， https://github.com/wbap/hackathon-2017-sample/wiki/
[5] Masahiro Miyata, Takashi Omori : Modeling emotion and inference as a value calculation system, BICA2017, 2017.
[6] Fabian Schoenfeld, Laurenz Wiskott: RatLab: an easy to use tool for place code simulations ， frontiers in Computational Neuroscience，(2013)
[7] Bendor, D. & Wilson, M. A.: Biasing the content of hippocampal replay during sleep，Nature Neuroscience，(2012)