NEURON AS AN AGENT

Shohei Ohsawa
The University of Tokyo
7 Chome-3-1 Hongo, Bunkyo, Tokyo
ohsawa@weblab.t.u-tokyo.ac.jp

ABSTRACT

The reason why swarm of agents solve real-world problem well is, interestingly, same as the principle of representation learning: good representation improves perforamance of machine learning. Most of the problem in real-world is not Markov decision process (MDP) but partially observed MDP (POMDP). On the POMDP environment, good observation yields good action. In this paper, we optimise a deep neural network as a multi-agent system as a natural extension from representation learning to multi-agent reinforcement learning for POMDP. To achieve that, we propose a novel learning framework, neuron as an agent (NaaA). In NaaA, an individual unit is considered as an agent, and they maximizing profit instead of minimizing error. To prevent dillemma, we borrow idea from machanism design, ka field of game theory. To this end, we show all the unit have valid price reflecting their contribution to performance at convergence. We confirm the result by numerical experiment using Atari and VizDoom.

1 INTRODUCTION

現在成功している深層強化学習のモデルの多くは、単一のエージェントが環境の観測から認知、行動決定といった一連のプロセスを担う。DQN (Mnih et al., 2015) や A3C (Mnih et al., 2016) は、Atari のスクリーン系列から最適な行動を決定する。AlphaGo (Silver et al., 2016) は囲碁の盤面から勝利に最も近い一手を選ぶ。DDPG (Lillicrap et al., 2015) は物理空間において、最適な筋肉の動きを決定する生物の動きを学習する。単一のエージェントを用いて強化学習を解くという試みは、人間の持つ身体性のアナロジーから考えると一見して妥当であるように思える。

しかし、現実世界の問題の多くはマルチエージェントによって解決される。すなわち、観測、認知、行動決定がすべて異なるエージェントによって行われる。各エージェントがセンシングした状態空間は、エージェント同士のコミュニケーションを通して統合されていき、最終的に報酬系との接点を持つアクチュエーターの行動に用いられる。たとえば、自動運転車は、自身だけでなく、周辺の車から得られた情報を統合することで衝突を減らすことができる。また、株のトレーダーは、多くの人物から提供された情報をもとに推論し、有用な情報提供者には対価を支払う。他にも、エレベーター制御 (Crites & Barto, 1998)、センサーネットワーク (Fox et al., 2000)、ロボットサッカー (Stone & Veloso, 1998) などがマルチエージェントによって解かれている。

マルチエージェントによる問題解決が現実の課題に対して有効である理由は、興味深いことに表現学習の原理と一致する：有用な表現は機械学習の性能を向上させる。現実にある問題の多くは、DQN や A3C が前提としているマルコフ決定過程 (MDP) ではなく部分的観測マルコフ決定過程 (POMDP) である (Sutton & Barto, 1998, p.258)。POMDP において真の状態は完全には不可視であり、良質な観測結果が良質な行動に結びつくため、アテンション、すなわち、有限のリソースを使って何を観測するかの設計が必要である。したがって、観測に用いられるエージェントが多いほど、将来的に獲得できる報酬も高くなる。また、複数のモデルを組み合わせることによって精度が高まるアンサンブル法としての側面が存在する。

本論文は、表現学習の群強化学習への自然な拡張として、ディープニューラルネットワークをマルチエージェントシステムとして考える。マルチエージェントによる観測および認知の枠組みは、センサー処理の分野ではエッジコンピューティング (Bonomi et al., 2012) としても知られている。エッジコンピューティングは分散環境を前提とした信号処理のモデルであり、一

つの処理系がすべてのデータを処理するのではなく、複数のセンサーの情報を一つのエッジサーバが集約し、複数のエッジサーバが次元削減したデータをデータセンターに送るという階層的な構造をしている。実際、人間の脳を観察しても、各神経細胞は独立して動作する。神経細胞は胚細胞からの発達段階において Nerve Growth Factor、BDNF といった神経栄養因子 (NTF) を追求することが知られており、十分な NTF を受け取ることができなかったニューロンはアポトーシスを引き起こして自死することが知られている (Almeida et al., 2005)。こうした観察から、各神経細胞を独立した生物として捉える見方はニューラルダーウィズム (Edelman, 1987) と呼ばれる。

マルチエージェントによる問題解決では、信頼度割り当て問題 (credit assignment problem) が発生する。既存研究として、複数のエージェントを想定したものがあるが (Agogino & Tumer, 2006)、エージェント間が直列に情報を伝達し合う状況における報酬分配については解が与えられていない。本論文で提案する学習の枠組み Neuron as an Agent (NaaA) は、ゲーム理論を想定し、エージェント間で報酬の分配を行う「通貨」を想定することで信頼度割り当ての問題を扱う。利益は報酬とコストの差分であり、ニューロンは下位のレイヤーから情報を買い、上位のレイヤーに対して信号を”売る”ことで報酬を得る。利益は、ニューロン自体が生成した付加価値であり、売価と買価を差し引いて計算される。

ゲーム理論を用いて機械学習の最適化を行う手法はいくつか提案されている (Stone & Veloso, 1998; Goodfellow et al., 2014) が、ゲームの設計方法によっては、ナッシュ均衡がパレート最適と一致しないというジレンマ (Holt, 2007) が発生する。ナッシュ均衡とは個々のエージェントによる報酬追及の結果として得られる帰結であり、パレート最適とは設計者意図する全体最適な帰結である。具体的に、個々のユニットが自身のコストを最小化した結果、入力ユニットに十分な報酬がいきわたらなくなる。

NaaA では、経済学におけるメカニズムデザインを用いて、分散環境における信頼度割り当ての問題を解く。メカニズムデザインでは、パレート最適な結果がナッシュ均衡の解と一致するようにする対戦略性の高い仕組みである。NaaA は、ユニットが複数財オークションを実行することによって最適な価格を決定する仕組みである。本論文では、メカニズムデザインの一つであるこの仕組みがエージェント間の社会的ジレンマを解決し、各ユニットが、他のユニットの本源的価値に基づく最適な価格設定を行うことを示す。具体的に、同じレイヤーに属するニューロン同士を競合学習させることで、neural agent が他のニューロンの価値を正直に申告することを示す。この報酬モデルは、エージェントがどの情報に注目すべきかというヒントを与える。すなわち、これはアテンションモデル (Xu et al., 2015) を拡張であると解釈できる。

実験では、標準的な強化学習のタスクによる数値実験を用いて NaaA が POMDP の問題の精度を高めることを示す。具体的に、Atari および VisDoom における環境を用いて、既存研究が DQN や A3C を上回ることを示す。

2 Related Work

深層強化学習を POMDP 環境に適用する場合、観測困難な環境からいかに真の状態を推定するかが重要になる。Deep Recurrent Q-Network (DRQN) (Sorokin et al., 2015) は、隠れマルコフ連鎖を想定し、リカレントニューラルネットワーク (RNN) を用いて真の状態を推定している。

マルチエージェントで強化学習の問題を解決する場合には、信頼度割り当て問題の解決が重要になる。そこで、エージェントの信頼度を、そのエージェントがいた場合と、いなかったと仮定した場合の差として定量化する研究が行われている。QUICR-learning (Agogino & Tumer, 2006) では、エージェント i が reward $R(a_t)$ の代わりに、そのエージェントがある行動 a_{ti} をとった場合 a_t と取らなかった場合 $a_t - a_{ti}$ の差、counterfractual reward $R(a_t) - R(a_t - a_{ti})$ の減衰和を最大化している。COMA (Foerster et al., 2017) は、actor-critic において critic が共通しており、actor がマルチエージェントであるという actor-critic の仕組みを考え、それぞれの actor が counterfractual reward を最大化するような仕組みを考えている。

これらの研究の問題は、情報がすべて共有されているという前提に立っており、信頼度を割り当てるという性善説に基づいている点にある。そのため、裏切ることが考えられる人物がいると、予想外の内容が学習されてしまう。たとえば、IoT のような実環境における問題を考えると、センサーを持っている主体は異なる人物であるために、協力行動をとるとは考えにくい。

本研究では、すべてのニューロンをエージェントとみなす方法を提案している。

3 NEURON AS AN AGENT

本セクションでは、ニューラルネットワークの構造について復習 (recap) した後、報酬分配のフレームワークである profit maximization について述べ、Valuation Net について説明を行う。

神経回路に含まれるニューロンは一つの細胞であるため、エネルギーを消費する。通常の細胞と同様に酸素や ATP がエネルギー源となり、これらはニューロンと接続した、アストロサイトから供給される。アストロサイトは脳の構造を支えるグリア細胞の一種であり、血管からニューロンへの栄養供給を行う。エネルギー量は有限であるため、不要なニューロンはアポトーシスによって死滅する。アポトーシスは NGF, BDNF などの神経栄養因子 (NTF) によって制御されるため、より多くの NTF を獲得できたニューロンが生存する。

こうした NTF による仕組みを選択圧としてとらえるニューラルダーウィニズムという考え方がある。ニューラルダーウィニズムでは、一つのニューロンを、バクテリアのような生物学における個体であるかのように扱う。環境に適応したニューロンが生存し、そうでないニューロンが死滅することで、有用なニューロンが生き残るようにする。

NaaA では、こうしたニューロンによる栄養追及の枠組みを強化学習でモデリングする。強化学習はエージェントと環境の相互作用を扱う。エージェントは環境に対して一つの状態を持ち、環境に対してアクションを行うことで状態遷移を行い、報酬を受け取る。アクションはエージェントの持つ状態とアクションの間の確率分布、方策を持つ。強化学習の目的は、長期的なエージェントの受け取る報酬の未来にわたる減衰和を最大化することである。NaaA では、ニューロンを一つのエージェントとみなし、環境を接続している他のニューロンと仮定する。

エージェント間の通信は一つのマシン内で行われてもよいし、マシン間で行われてもよい。

ニューラルネットワークにはパーセプトロン、ボルツマンマシンなど様々なものが存在するが、NaaA が対象とするニューラルネットワークの種類は、多層パーセプトロン、CNN、RNN、LSTM のような、微分可能 (differentiable) ニューラルネットワークである。これは、出力をパラメータで微分可能な関数群で構成される。直感的には、2017 年現在の TensorFlow でサポートされるものすべてを指す。一方で、ボルツマンマシンのような確率ベースのものは今回は対象としない。

微分可能ニューラルネットワークは、本来は人間の神経回路を模して造られたものであり、神経を構成するニューロンをユニットという形で抽象化している。ユニットは互いにつながっており、これはニューロンの軸索に対応付けられる。ユニットは接続されたユニットから受け取った信号に対して反応し、信号を出力する。

ユニットの計算方法は、既存のニューラルネットワークとほとんど同じである。ユニットは、一つのスカラー値を信号として出力する。この信号は、他のユニットから受けた入力に基づき計算される。線形演算、すなわち重み付き和を計算するユニットもあれば、ReLU のような活性を計算するユニットも存在する。

ユニットを一つのエージェントとみなすことで、最適化する対象の問題はマルチエージェント強化学習とみなすことができる。マルチエージェント強化学習における問題とは、環境が与えた報酬 R_0 を複数のエージェント間でどのように分配するかにある。

NaaA は、N 個のエージェントで構成されるマルチエージェントシステムである。各ユニットは本来の性質に加えて、以下の性質を持つ。

- N1: （利己性）エージェントは、系全体の期待リターン G_{0t} ではなく、自身の期待リターン G_{it} の最大化を目的として行動する。
- N2: （報酬の分配）各エージェントが受け取る報酬 R_i の総和は、マルチエージェントシステム全体が外的環境から得る報酬 R_0 に等しい。
- N3: （NOOP）エージェントは、期待リターンが 0 の NOOP (no operation) という行動をオプションとして持つ。NOOP では、エージェントは何も入力せず、何も出力しない。
- N4: （コスト）エージェントには、生存コスト α_i が存在し、タイムステップごとに負の報酬として消費される。NOOP の状態では、生存コストは消費されない。
- N5: （取引）エージェント間の信号の送受信は取引である。すなわち、エージェントは信号を他のエージェントに伝達する際に、信号と引き換えに報酬を受け取る。

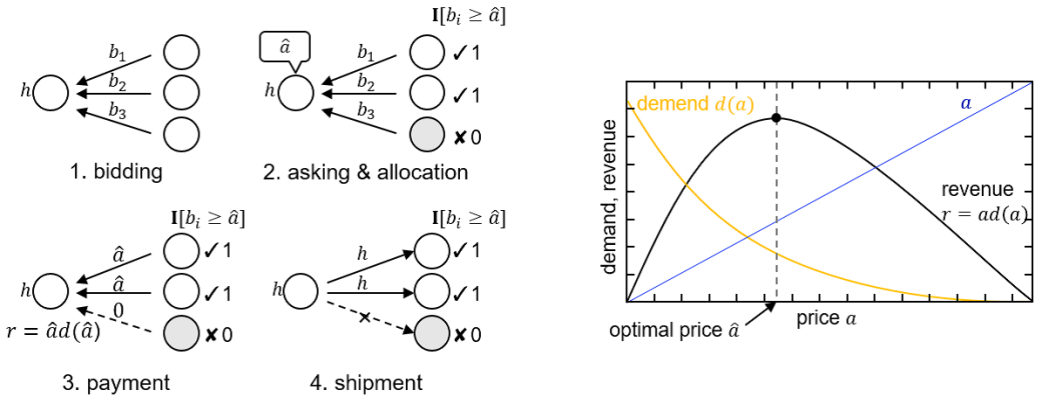

Figure 1: 本研究では、digital goods auction の枠組みを用いて、報酬の最適配分を実現する。

N6: （Envy-free 性）信号を同時に複数のエージェントが購入する場合、それはすべて同じ価格で取引される。すなわち、エージェントは同じ時間 (timing) に一つの価格 (price) を持つ。

3 つめの仮定は、ニューロンのアポトーシスに相当する。NOOP が選択されるのは、それ以外のすべての行動の期待報酬が負であった場合である。

3.1 Profit Maximization Framework

これまでのニューラルネットワークの目的は、予測誤差 e にあった。そのため、各ユニットはバックプロパゲーションによって $\partial e/\partial y$ を受け取り、e が小さくなる方向に y の大きさを制御していた。NaaA において、エージェントの目的は大域的誤差の最小化ではなく、エージェント自身の利益の最大化である。ここで、利益とは、エージェントが受け取る売上と、支払うコストの差である。すなわち、各エージェントは次の関数を最大化する。

$$R_t = r_t - c_t \tag{1}$$

エージェントの目的関数は次のように表現される。

$$\mathbb{E}_\pi \left[\sum_{i=0}^{T} \gamma^i (r_{t+i} - c_{t+i}) \right] = \mathbb{E}_\pi \left[\sum_{i=0}^{T} \gamma^i r_{t+i} \right] - \mathbb{E}_\pi \left[\sum_{i=0}^{T} \gamma^i c_{t+i} \right] \tag{2}$$

$$= G_t^{\text{in}} - G_t^{\text{out}} \tag{3}$$

ただし、$G_t^{\text{in}} \equiv \mathbb{E}_\pi \left[\sum_{i=0}^{T} \gamma^i r_{t+i} \right]$、$G_t^{\text{out}} \equiv \mathbb{E}_\pi \left[\sum_{i=0}^{T} \gamma^i c_{t+i} \right]$ である。この式は、エージェントが終端状態までに生み出した付加価値に等しい。これを行うために、エージェントは売上の最大化とコストの最小化を同時に行うことがわかる。

この式を通常の枠組みに沿って解くと、コスト最小化によってエージェントは他のエージェントに対して金額を支払わないことが正解となる。そのため、すべてのエージェントがフリーライダーとなり、全体としての報酬が下がるという、パレート劣位問題が発生する。NaaA ではこれを回避するため、オークションの仕組みを用いて、同じユニットの出力先にあるユニットを競争させることによって最適価格を決定する。

説明を単純化するために、DQN のような l 層からなるフィードフォワードネットワークを考える。なお、任意の DAG 構造（時間方向に展開された RNN やホップフィールドネットワーク）でも同様のことを考えられるため一般性を失わない。各エージェントが受け取る報酬の総和はが環境が与える報酬に対して等しい必要がある。すなわち、次の式が成立する。

$$R_{S,t} = \sum_{i \in S} R_{it} \tag{4}$$

最適な分配方法としてはいくつか考えられるが、ここでは系全体からエージェントを取り除いた場合の期待損失 D_{it} にを R_{it} を近づけるという方法を取る**?**。この方法では、すべてのエージェントが独立に報酬に貢献している場合に、$\sum_{i \in S} D_{it} = R_{S,t}$ が成立する

今回は、制約条件として、報酬は信号同士の取引によって与えられるというものを加える。したがって、ニューロンは、そこで、アクチュエータ 0 と接続している複数のユニット集合 $F_0 \subset S$ を考える。この時、アクチュエータは得られた報酬のうち、次のようにして各ノードへの分配を行う。

$$R_{0t} = \sum_{i \in V_0} r_{it} \tag{5}$$

ただし、r_{it} は貢献度に応じた報酬額である。次に、ユニット i は計算に必要なデータを「購入」するために、接続されている（すなわち V_i に含まれる）他のユニットに対して、報酬を次のようにして分配することを考える。

$$c_{it} = \sum_{j \in V_i} r_{jt} + \alpha_{jt} \tag{6}$$

ただし、α_i は c_{it} が本質的に備えているコスト（観測コスト、計算コスト）である。この仕組みを再帰的に繰り返し、入力層まで繰り返すと、次の式が成立する。

$$R_{0t} = \sum_{i \in V_0} r_{it} = \sum_{i \in S} (r_{it} - c_{it}) \tag{7}$$

したがって、$R_{it} = r_{it} - c_{it}$ が成立する。この式は、R_{0t} は、各ユニットが生み出した付加価値 $r_{it} - c_{it}$ の合計として表現されることを意味している。

この付加価値のことを、NaaA では経済学のメタファーを用いて、利益 (profit) と呼ぶ。利益は強化学習において追求の対象となる報酬 (reward) に相当する。同様に、r_t を収益 (revenue)、c_t をコスト (cost) と呼ぶ。すなわち、NaaA の枠組みでは、各エージェントは利益の最大化を目的として、収益最大化とコスト最小化を同時に行うことになる。

通常のニューラルネットワークの枠組みを用いて最適化を行うと、すべてのエージェントが支払うコストは 0 に収束する。したがって、この最適化問題の自明なナッシュ均衡解として得られるのは、すべてのユニットに対する報酬が 0 であり、アクチュエーターのみが報酬を獲得するという状況である。これは認知や入力を軽視した従来の状況と合致する。

従来はこれでもよかったが、入力にコストが生じる場合、パレート劣位な状況が生じる。すなわち、入力ユニットに対して報酬が与えられないと、入力側のインセンティブがなくなり、$R_{it} = -\alpha_{it}$ となる。この場合、入力ユニットは観測をしない方がよいという判断になり、出力を停止する。その結果、アクチュエーター側では、行動の決定に必要なデータがいきわたらなくなり、報酬が低くなる。こうした状況を避けるために、期待報酬が負である場合、エージェントは行動しない。

これを避けるために、NaaA ではオークションの仕組みを用いて資源の最適配分を行う。

ここで、売上とコストの選定方法について分けて説明を行う。まず、売上について考える。エージェントは利益を最大化したいため、利益は次のようにして与えられる。

$$r_{it} = \max_{a \geq 0} a d_t(a) \tag{8}$$

ここで、a は価格、$d_t(a)$ はユニット i の信号に対する価値を a 以上と評価しているエージェントの数であり、需要 (demand) と呼ぶ。同様の式で、右辺を最大化する a を最適価格と呼び、$\hat{a}_{it}$ で表す。

次に、コストについて考える。エージェントが他のエージェントに対して支払うコストは、購入に成功した場合は売り手の提示価格 $\hat{a}_{it}$ を支払い、そうでない場合は支払うコストは 0 になる。コスト最小化の枠組みから考えると、支払うコストが 0 である方向を選ぶことがあるが、エージェントは常にデータを購入しない選択をする可能性がある。しかし、これは潜在的には将来の売り上げを棄損していることになる。データを購入せず、劣悪なデータを流した場合の売上の低下を考える。これは、正しいデータを購入した場合とそうでない場合の長期 r_{it} の差に等しい。すなわち、

$$o_t = \mathbb{E}_\pi \left[r_{i,t+1} \mid a_t = 1 \right] - \mathbb{E}_\pi \left[r_{i,t+1} \mid a_t = 0 \right] \tag{9}$$

$$= \mathbb{E}_\pi \left[R_{i,t+1} \mid a_t = 1 \right] - \mathbb{E}_\pi \left[R_{i,t+1} \mid a_t = 0 \right] \tag{10}$$

$$= Q(s_t, 1) - Q(s_t, 0), \tag{11}$$

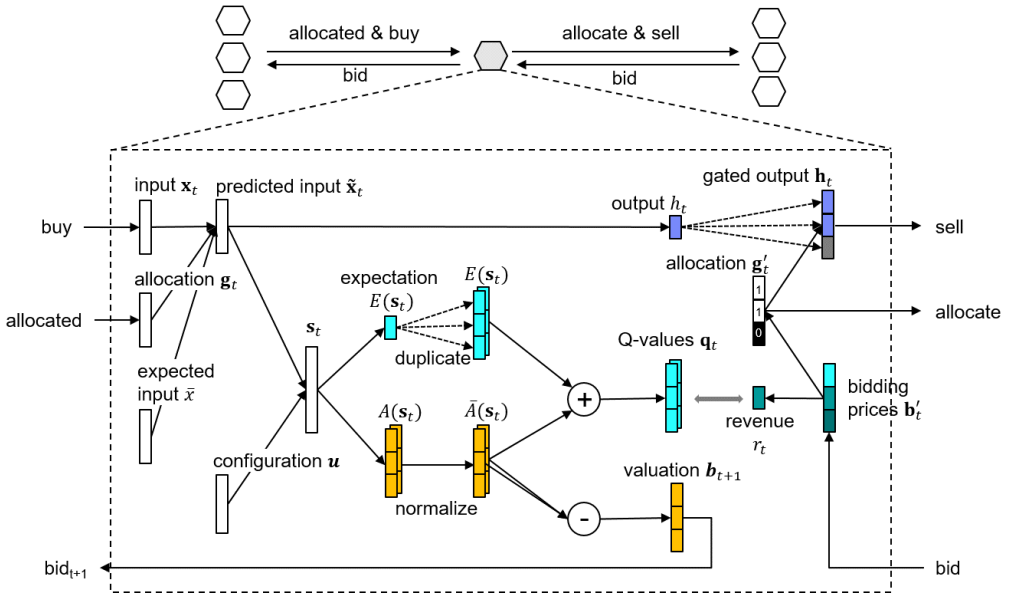

Figure 2: Valuationn Net は情報の価値を評価し、bidding price を決定する。下位のニューロンに対して入札し、信号を購入する。購入したデータを用いて、データを次のニューロンに対して売る。

ただし、Q は状態行動価値関数であり、一手先のコストはどちらの行動を選んでも一定であると仮定している。o_{it} を counterfractual state-action value と呼ぶ。これは QUICR (Agogino & Tumer, 2006) と導出は異なるが等価である。すなわち、エージェントが支払うコストは、データの購入に成功した場合は $\hat{a}_{it}$ であり、それ以外は o_{it} となる。

では、コストを最小化するためのエージェントの入札額 b_{it} は何か。これについては次の定理が成立する。

Theorem 3.1. コストを最小化する最適な入札額は $b_{it} = o_{it}$ である。

証明については Appendix を参照。

すなわち、エージェントは自身の機会損失のみを問題にすればよい (!) したがって、NaaA のメカニズムでは、エージェントはあたかも他のエージェントを価値評価 (valuation) し、その価値を正直に申告していることを意味する。

系として次の解が得られる。

Corollary 3.1. *The Nash equivalem of the envy-free game* $(\mathbf{b}, q)$ *is* $(\mathbf{o}_t, \max_q qd_{\mathbf{o}_t}(q))$.

3.2 Valuation Net

残る問題は、$\mathbf{o}_t$ をいかに推定するかである。この推定には様々な方法が存在しており、多くのメソッドを使うことができるが、本論文では Q の推定に Q-learning を採用する。ただし、SARSA や actor-critic などの on-policy な方法も使うことができることを補足する。

図 2 に示す Valuation Net は、通常のニューラルネットワークのユニットに、Q-learning による valuation を組み合わせたネットワークである。まず、上部はエージェント間の通信について示したものである。ニューラルネットワークではユニットを円で表現するのが通例であるが、ここではユニットをエージェントとしてみなすことを強調して、六角形で一つのユニットを表現している。エージェント間では、通常のニューラルネットワークと同様の信号の通信以外に、取引に関する通信 (allocate, buy, sell & bid) が発生する。

Valuation Net では、状態 $\mathbf{s}_t$ として、予測後入力 $\tilde{\mathbf{x}}_t$ および入力に依存しない構成情報 $\mathbf{u}$ を横につなげたベクトル $(\tilde{\mathbf{x}}_t^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}$ を用いる。構成情報の一例としてはユニットのパラメータがあげられ、たとえば重みやバイアスの情報を用いることができる。

状態からの Q 関数の予測にニューラルネットワークを用いる。エージェントが受け取った売上に基づき時間差分 (TD)-誤差 が計算され、ネットワークが訓練される。ネットワークの構成にはこれまでの deep Q-learning で用いられている二重化ネットワーク (dualing network) (Wang et al., 2015) のテクニックを用いる。オリジナルの文献 (Wang et al., 2015) で述べられている二重化ネットワークは、学習を加速するために、状態関数と、Q 関数との差分を別々に予測する手法である。Dosovitskiy & Koltun (2016) はこれに対して、差分の要素の挿話が 0 になるように正規化するよう改良している。本研究では Dosovitskiy & Koltun (2016) の手法に従い、期待値 $E(\mathbf{s}_t)$ と正規化差分 $\tilde{A}(\mathbf{s}_t)$ を別々に求める。

Q 関数は次のように表現される。

$$Q(\mathbf{s}_t, a_t) = E(\mathbf{s}_t) + \tilde{A}(\mathbf{s}_t, a_t) \tag{12}$$

$$\sum_{i+1}^{k} \tilde{A}_i(\mathbf{s}_t, a_t) = 0 \tag{13}$$

第 2 式を満たすために、まず、$\mathbf{s}_t$ に基づいた予測を行い、次のような正規化を行う。

$$\tilde{A}_i(\mathbf{s}_t, a_t) = A_i(\mathbf{s}_t, a_t) - \frac{1}{k}\sum_{j=1}^{k} A_j(\mathbf{s}_t, a_t) \tag{14}$$

次に、valuation を行い、bidding price $\mathbf{b}_t$ を求める。b_{it} の値は式 11 および式 12 より、最適な入札価格 $\hat{b}_{it}$ は次のように計算できる。

$$\hat{b}_{it} = \tilde{A}(\mathbf{s}_t, 1) - \tilde{A}(\mathbf{s}_t, 0) \tag{15}$$

Valuation Net ではこの式に基づき、advantage の出力を引き算することで入札価格を計算している。

4 Experiment

5 Discussion

6 Conclusion and Future Works

本論文では、POMDP の問題設定において良質な特徴表現を得るために、ニューラルネットワーク上の各ユニットをエージェントとして扱うフレームワーク、NaaA について述べた。NaaA のフレームワークでは、ジレンマ問題を解決し、それぞれのエージェントの持つ付加価値がナッシュ均衡として得られ、全体としてパレート最適になることを示した。入札価格の決定アルゴリズムの一つとして、Q-learning に基づくネットワーク Valuation Net を示した。評価実験では、Atari と VizDoom を用いた実験を行い、実験結果が既存手法よりもよくなることを示した。

今後の方向性として、高速化、Valuation Net を A3C などの on-policy な手法で置き換えるといった方向性の他、神経科学的な説明を可能にしていくといった方法、遺伝的アルゴリズムとの組み合わせが挙げられる。

Appendix

A.1 定理 3.1 の証明

買い手の獲得する生涯報酬 G は次で与えられる。

$$G(b, q) = g(b, q) \cdot (v - q) + G_0, \tag{16}$$

ただし、g は割当 (allocation) であり、G_0 はユニットを購入しなかった場合の生涯報酬である。割当は、$g(b, q) = H(b - q)$ が成立する。H はステップ関数である。

買い手にとって売り手が提示する asking price q は未知であるため、q を台 $[0, \infty)$ の上の確率変数であるとして扱い、期待値 $\mathbb{E}_q[G(b,q)]$ を最大化することを考える。このとき、

$$\begin{aligned}
\frac{\partial}{\partial b}\mathbb{E}_q[G(b,q)] &= \frac{\partial}{\partial b}\int_0^\infty (H(b-q)\cdot(v-q)+G_0)p(q)dq \\
&= \frac{\partial}{\partial b}\left[\int_0^b (v-q)p(q)dq + G_0\int_0^\infty p(q)dq\right] \\
&= \frac{\partial}{\partial b}\int_0^b (v-q)p(q)dq \\
&= (v-b)p(q=b)
\end{aligned}$$

したがって、 $\mathbb{E}_q[G(b,q)]$ が最大となるための条件は $b = v$ である。

REFERENCES

Adrian K. Agogino and Kagan Tumer. Quicr-learning for multi-agent coordination. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pp. 1438–1443. AAAI Press, 2006. ISBN 978-1-57735-281-5. URL http://dl.acm.org/citation.cfm?id=1597348.1597417.

RD Almeida, BJ Manadas, CV Melo, JR Gomes, CS Mendes, MM Graos, RF Carvalho, AP Carvalho, and CB Duarte. Neuroprotection by bdnf against glutamate-induced apoptotic cell death is mediated by erk and pi3-kinase pathways. *Cell death and differentiation*, 12(10):1329, 2005.

Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pp. 13–16. ACM, 2012.

Robert H Crites and Andrew G Barto. Elevator group control using multiple reinforcement learning agents. *Machine learning*, 33(2):235–262, 1998.

Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*, 2016.

Gerald M Edelman. *Neural Darwinism: The theory of neuronal group selection.* Basic books, 1987.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926*, 2017.

Dieter Fox, Wolfram Burgard, Hannes Kruppa, and Sebastian Thrun. A probabilistic approach to collaborative multi-robot localization. *Autonomous robots*, 8(3):325–344, 2000.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Charles A Holt. *Markets, games, & strategic behavior*. Pearson Addison Wesley Boston, MA, 2007.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Ivan Sorokin, Alexey Seleznev, Mikhail Pavlov, Aleksandr Fedorov, and Anastasiia Ignateva. Deep attention recurrent q-network. *arXiv preprint arXiv:1512.01693*, 2015.

Peter Stone and Manuela Veloso. Towards collaborative and adversarial learning: A case study in robotic soccer. *International Journal of Human-Computer Studies*, 48(1):83–104, 1998.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of The 32th International Conference on Machine Learning (ICML-15)*, 2015.