画像認識による言語の対応付けを利用した強化学習

飯田 啓太[1,a)]　鶴岡 慶雅[2]

概要：自然言語理解は人工知能分野の研究テーマのひとつであり、その一環として行動や物体に対して言語を対応づけることを目指すものがある。従来のルールベースによる手法に代わるものとして近年では強化学習が利用されるようになってきており、そのテストのための環境として頻繁にゲームが用いられている。ゲームによるシミュレーション環境で行なった学習でも高い性能を発揮しているが、その一方で学習には多くの試行が必要であるという欠点が存在していた。そこで、本研究では一般的な事前知識を加えることで学習に要する試行を減らすことを目標とする。事前知識として画像認識を活用し、これによって学習に要する試行回数がどのように変化するかについて実験、調査を行なった。

Reinforcement Learning Leveraging Grounded Language by Image Recognition

Keita Iida[1,a)]　Yoshimasa Tsuruoka[2]

Abstract: Grounding natural language is an important task in artificial intelligence research. Recently, reinforcement learning has been applied to various tasks, and games are frequently used as testbeds. It has performed very well, but it requires a large number of trials to learn. The purpose of this research is to decrease the number by using prior knowledge. We investigated how leveraging image recognition as knowledge affects the number of trials.

1. はじめに

1.1 背景

ゲーム AI の研究は、ルールが明確であり実験がコンピュータ内で完結させられることや、現実世界のコンパクトなシミュレーション を行うことができることなどから人工知能分野の発展に有用である [1]。

ゲーム AI は当初プログラマがゲームのルールなどに基づいてゲームごとに作成する手法が主流であったが、近年では強化学習およびその応用が注目を集めている。基本的な強化学習の手法である Q 学習 [2] にニューラルネットワークを組み合わせた Deep Q-Network という手法は、ゲームの事前知識が無い状態から画像の入力のみで学習を行い高い性能を発揮した [3]。

しかし、実際に人間がゲームをする際には画面に表示された画像のみではなくさまざまな情報を利用している。人間の用いる情報の一つとしては言語情報が挙げられる。画面に表示されるスコアなどの文字情報から外部マニュアルに至るまでその形態は多岐にわたり、コンピュータがこのような情報を利用できればより高度なタスクにも対応できるようになると考えられる。実際に人工知能の研究の一環としてコンピュータに人間の用いる自然言語を理解させる試みは古くから行われてきており、その中でも言語と意味の対応付けの問題は長きに渡って研究されてきた。当初は Winograd [4] が、自然言語によって記述された指示をコンピュータがシミュレーションの中で実行できることを示したことで注目された分野であったが、この時の手法が考えられる入力を全てコーディングするというものであったため、近年では言語と意味の対応を学習するという手法

[1] 東京大学工学部電子情報工学科
Department of Information and Communication Engineering, The University of Tokyo
[2] 東京大学大学院情報理工学系研究科電子情報学専攻
Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo
[a)] iida@logos.t.u-tokyo.ac.jp

がとられるようになった。これにより、膨大な数のある入力ケースに対しても対応することが期待されている。本分野における過去の研究としては、Branavan ら [5] による、ゲームマニュアルの利用方法を強化学習によって学習し、マニュアルを利用しない場合と比較して AI の性能が有意に向上することを示した研究や、Hermann ら [6] による自然言語で記述された指示を実行するという課題に対し、単語レベルでの意味理解を行って学習できることを示した研究が挙げられる。だが、これらの手法が対応できるのは過去に経験した単語や表現のみであり、その学習には相応の時間を要する。そして、それにもかかわらず未知の言語表現や環境に対しては新たにコストをかけて再度学習する必要があるため、学習ごとのコストが大きい場合には適応が難しいことが課題となっている。

1.2 目的

本研究は、強化学習を行う場合には多くの試行が必要であるにもかかわらず未知の環境に対応しにくいという問題を踏まえ、言語理解を行うことにより新たな環境への対応力を上げること、またこの言語理解の助けとして画像認識による解析を利用し、既存の知識として活用することで、学習の試行回数を減らすこと、言語と意味の対応を強化することを目的とする。

1.3 本稿の構成

本稿では、第 2 章で本研究に要する知識や関連する研究を紹介し、第 3 章では本研究における提案手法を示す。第 4 章では提案手法の検証のため実験として行なった学習について述べ、第 5 章で本稿のまとめと今後の課題について述べる。

2. 関連研究

2.1 強化学習

機械学習の手法の一つとして現在盛んに研究されているものに強化学習がある。強化学習は、エージェントが現在の状態をもとに選択肢から行動を一つ選択、それに応じた報酬を受取れるような環境の中で、最終的に受け取る報酬の総和が最も大きくなる方策を学習するというものである。

機械学習の手法としては他に教師あり学習が挙げられる。教師あり学習と比べて強化学習では、状態に対する評価などの教師データが存在せず、報酬を最大化するような方策を試行錯誤して発見しなければならないため学習にかかる時間は大きくなるが、入力に対する教師データを用意する必要がないという点で異なっている。

2.2 Q 学習

強化学習の代表的な手法に Q 学習がある [2]。Q 学習は有限マルコフ決定過程 (Markov decision process, MDP) においては学習の収束することが保証されているため、問題を MDP としてモデル化できる場合に用いられる手法である。MDP は状態の集合 $S = \{s_1, s_2, ...\}$、行動の集合 $A = \{a_1, a_2, ...\}$、遷移関数 $P(s'|s, a)$、報酬関数 $R(s)$ の 4 要素で表され、Q 学習では S と A の要素数が有限なものを扱う。Q 学習では現在の状態 $s \in S$ に対して期待報酬を最大化するような行動を行う方策を学習する。そこで、方策として状態 s で行動 a をとるものを $\pi(s) = a$ とするとき、目的関数を

$$Q^{\pi}(s) = E[\sum_{t}^{\infty} \gamma^t R(s_t)|\pi, s_0]$$

として設定する。ここで γ は割引率という定数であり、近い将来の報酬を重視することで不要な行動を行わないための値である。この関数は最適な方策 π^* においてはベルマン方程式として知られる

$$Q^{\pi^*}(s) = R(s)|_{\pi^*} + \gamma \sum_{s'} P(s'|s, a) Q^{\pi^*}(s')$$

の形で表され、これに関数 Q を近づけていくことで学習を行う。基本的な Q 学習では学習率 α を用いて、以下の式で更新を行う。

$$Q(s, a) = Q(s, a) + \alpha(R(s, a) + \gamma \max_{a} Q(s', a) - Q(s, a))$$

多くの場合、定数は $0.9 < \gamma < 1$、 $0 < \alpha < 0.1$ 程度の値に設定される。

2.3 畳み込みニューラルネットワーク

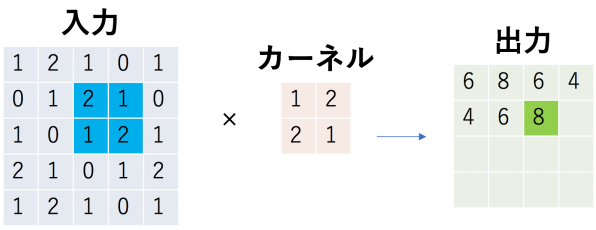

図 1 畳み込みニューラルネットワークの動作例

畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) はニューラルネットワークのうち、畳み込み層やプーリング層を持つものであり、全結合層のみでは処理が大きくなる画像などの解析に適している [7]。畳み込み層などで入力の大きさを減らしたのちに全結合層に結果を渡すことが多く、学習を軽量化するために用いられる。

畳み込み層ではカーネルと呼ばれる小さな領域を移動しながら乗算することで処理を行う。この操作を行う様子を図 1 に示す。ここでは入力サイズを 5×5、カーネルのサイズを 2×2、カーネルの移動距離 (ストライド) を 1 としている。このカーネルのパラメータを変化させることで学

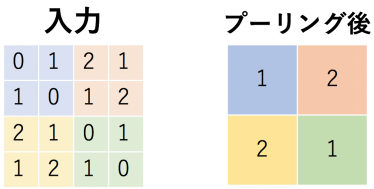

図 2　プーリングの例

習を行う。

プーリング層は主に入力の圧縮を行う層である。この操作は計算のコストを下げることや出力の安定化を主な目的としており、複数の領域の最大値を返す、平均値をとるなどの手法が取られる。最大値プーリングを行なった時の様子を図 2 に示す。ここでは 4×4 の入力において、2×2 の領域に最大値プーリングを行なっている。カーネルの大きさや移動距離、プーリングの数などは実験中に固定されるため、これらのパラメータはあらかじめ決めておく必要がある。

画像の処理に CNN を用いることで画像内の位置情報を保持しながら小さい計算量でサイズを下げることができる。これは、全結合層と比べて CNN では限られた領域に対してのみ演算を行うためであり、周囲の情報のみで計算されるため、画像内での位置が出力にそのまま影響を与えるためである。

2.4　画像認識

画像認識は画像や動画から何らかのオブジェクトを検出するという技術である。人間にとっては日常的に行っている作業であり非常に容易である一方、コンピュータで実行することは難しく、現在でもコンピュータビジョンの分野で研究が行われている。従来はテンプレートマッチングなどの手法が用いられていたが、ニューラルネットワークの発展に伴い CNN を利用した手法が提案された [7]。このような CNN を用いて画像認識を行う手法に YOLOv3 がある [8], [9]。これはオブジェクトの検出からクラス分類までの全てを一つの CNN で行うという手法で、画像を一度ニューラルネットワークに通すだけで高速に検出が行えるため、リアルタイムの画像認識にも用いられている。

2.5　言語と意味の対応

言語とその意味の対応付けという課題は自然言語の研究分野の一つであり、様々な方法での解決が試みられてきた。近年では強化学習を応用することで言語と行動や物体の対応を学習するという手法がとられ、人間が直接コーディングすることなく対応付けをする研究が進んでいる。特定の環境に学習を行うエージェントを配置し、そこに自然言語で記述された指示を与え、指示の通りに行動できた場合に報酬を与えるといった方法で強化学習を実行しており、この際指示や観測した環境をニューラルネットワークに通すことで行動を決定する。すなわち、Q 関数の値をニューラルネットワークに出力させ、Q 学習を行うことで対応付けを行うことが多い。この方法では環境とエージェントを用意するだけで自動的に対応付けを行うことができるが、強化学習のデメリットを引き継いでおり、学習に時間がかかることと未知の入力に弱いことが欠点である。

2.6　画像からの言語理解

実際に言語の意味と物体および行動の対応付けを行う例として、Hermann ら [6] による研究がある。この研究では 3D シミュレーションゲーム環境下で言語を学習する強化学習を行うというもので、ゲームの空間内に配置されたオブジェクトのうち、指示されたオブジェクトまで移動を行うというものである。ここではゲームの画面と指示オブジェクトを表す文章をニューラルネットワークへ入力し、その出力からとるべき行動を決定するというモデルを用いている。さらにこの学習で得られたモデルを用いて新たな入力に対する学習を行うと、はじめから学習を行なった場合と比べて早く学習が行われることが判明している。

3.　提案手法

本研究では言語情報とその意味および行動の対応を学習するという課題において、すでに存在するデータを事前知識として用いることで学習の回数を減らすことが目標である。実験環境としては簡単なゲームをシミュレーションとして用い、事前知識としては既存の画像認識ライブラリを利用する。本研究で提案する手法はゲームの画像から画像認識によってゲーム内のオブジェクトを抽出し、その情報を利用することで学習の回数を減らすものである。例えば、「猫に触る」という動作を実行するためには「猫」と「触る」を共に理解する必要があるが、ここで画像認識を用いることで画面から「猫」を検出できれば、「触る」の部分を理解すれば良い。また、画像認識を用いるため未知の動作対象にも対応できる可能性が高い。例えば「犬を触る」という動作において「犬」を知らない状況であっても「触る」を理解していれば「犬」を検出することで動作が正しく実行できることが見込まれる。そこで、言語と意味の対応付けにおいて強化学習を行う際に画像と並列してオブジェクト検出の結果、すなわち画像のどの部分に何があるかを入力することで情報を追加する。理想的には画像認識の検出できる範囲内であれば未知のものであっても対応できると期待される。

4.　実験

4.1　概要

本研究では言語の対応付けの学習において、事前知識として画像認識を活用することがどのような効果をもたらすかゲームによるシミュレーションを用いて確認した。具体

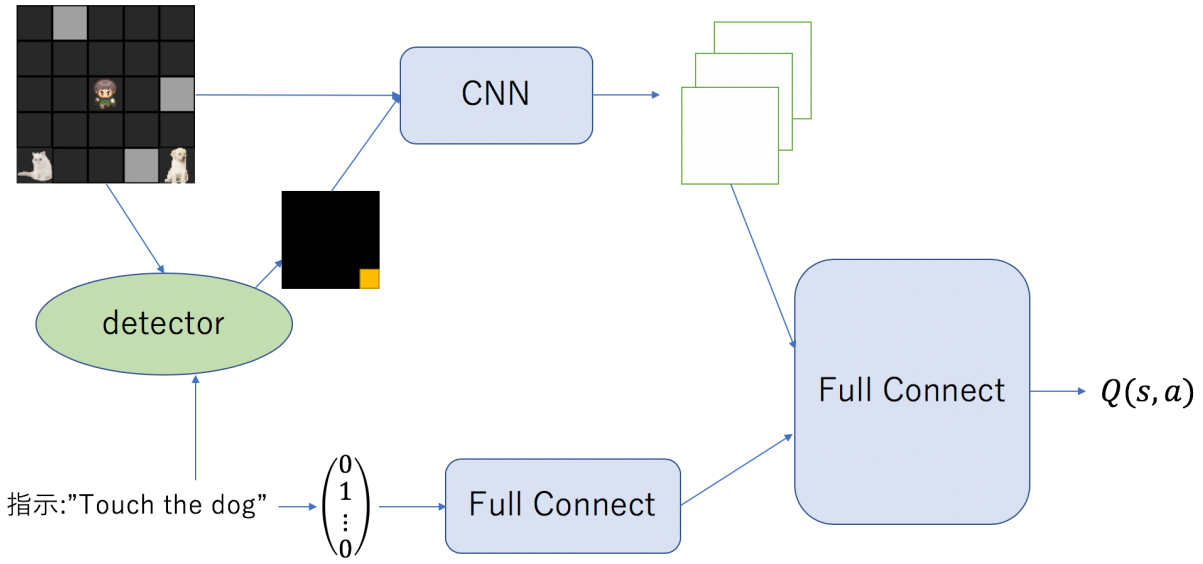

図 3　提案モデル

的には、事前知識の有無が学習の速度に与える影響を実験により調査した。

4.2　実験内容

図 4　ゲーム画面

本研究では以下のシミュレーション環境において学習を行った。

- 5 × 5 のグリッド平面上をプレイヤーは移動する。以下このグリッドそれぞれをマスと呼ぶ。
- プレイヤーは 1 回の移動で上下左右のいずれかに 1 マス分動くことができる。
- 平面上にはオブジェクトの置かれたマスが存在する。

今回の学習では自然言語で記述された指示の達成を目標とした。具体的な設定を以下に示す。

- ランダムにオブジェクトが配置される。オブジェクトはランダムに 2 種類が一つずつ重ならないように配置される。
- オブジェクトは全 5 種類の画像から選択される。
- “Touch xxx”の形で特定のオブジェクトに触れる指示が出される。
- 指示の対象はグリッド平面上のオブジェクトからランダムに選ばれる。
- 一定回数以内の移動で指示を達成できた場合に正の報酬が得られる。今回は 20 回を移動回数の上限とした。
- 指示対象でないオブジェクトに触れた場合には負の報酬が得られる。

実際の画面の一例を図 4 に示す。プレイヤーは与えられた指示とゲームの画面をもとに行動を決定する必要があり、画面と指示以外は観測できない。また画面内に存在するオブジェクトが同じであってもゲームのクリア条件となる指示の対象はランダムで決まるため、画面情報のみでは動作の対象が決定できず、クリア率を上げることが難しいゲームとなっている。このため、ゲームを高い割合でクリアするためには指示の理解が不可欠である。

学習のモデルは図 3 に示すものを用いた。画面の入力には画面の色情報の 2 次元配列、すなわち画像サイズと同じ 160 × 160 の大きさを持ち、各ピクセルの RGB 値を要素に持つものを用いた。この際、RGB 値は 0 以上 1 以下の範囲に変換した。画像認識には重みを学習済みの YOLOv3 を用い、指示対象のオブジェクトが存在すると判定されたピクセル領域に YOLOv3 の出力した確度を、それ以外の領域には 0 を格納したものを利用した。対象のオブジェクトが検出されなかった場合には全ての領域に 0 を格納した。CNN には画面情報の 2 次元配列 3 つに画像認識結果の 2 次元配列を合わせた、合計 4 チャンネルの画像を入力した。言語情報には指示を Bag of Words の形式に変形させたものを用いた。つまり指示に含まれる単語に対応する

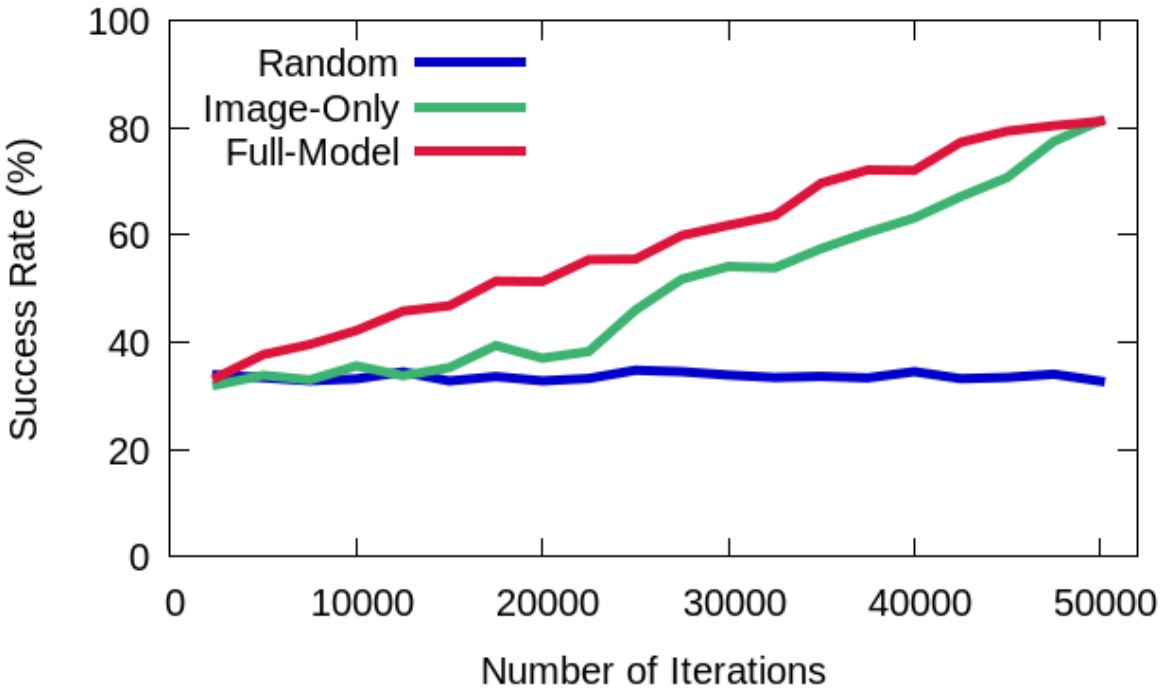

図 5 ゲームの実行回数に対する各モデルの指示達成率

次元のみを 1、それ以外の次元を 0 としたベクトルに変換し入力とした。学習では ϵ-greedy で行い、ϵ の値は学習の進行に合わせて 1 から 0.1 まで線形に減少させた。また今回の実験のベースラインとしてはランダムに行動するプレイヤー、図 3 のモデルから画像認識を利用する部分を取り除いたプレイヤーの 2 つを用意した。後者は CNN への入力において画像の RGB 値のみを用いたものであり、それ以外の構成は図 3 のものと等しい。実験では各モデルについて 50,000 回のゲームを実行し、学習を行った。その上で 2,500 回ごとに分割し、その区間内でゲームクリアの割合を測定した。ここでゲームクリアとは、一定回数以内に指示を達成することを指す。

4.3 結果

図 5 に今回の予備実験の結果を示す。3 種類のエージェントについて、図 3 のモデルを用いたもの (Full-Model)、図 3 のモデルから画像認識によって得られる情報を欠落させたもの (Image-Only)、ランダムに行動するもの (Random) それぞれの指示達成率をグラフにしたものであり、横軸がゲームの実行回数、縦軸がクリアしたゲームの割合を示している。

まず Random を見ると完全にランダムな動きを行なった場合のゲームクリア率が約 1/3 であることがわかる。今回の設定では移動回数に制限があるため、いずれのオブジェクトにもたどり着けない場合が発生した結果ゲームのクリア率が半分を下回っている。Image-Only については適切に学習が行われており、初期では Random と同程度の性能であるが、試行回数が増えるにつれゲームのクリア率が改善され、言語とゲームの対応が行われていることが読み取れる。このことから、今回の課題では画像の情報のみでも学習が十分可能であることがわかる。一方 Full-Model では Image-Only と比べてさらに高い達成率となっていることが確認できる。Image-Only が学習の初期に停滞しているのに対して Full-Model は学習の早い段階から高い性能を示しており、オブジェクトの検出によって動作対象を割り出すことが言語とゲームの対応を学習を助けることができると考えられる。

5. 終わりに

本研究では学習において有用な知識を用いることで、少ない試行回数であっても学習が早く進行することが示された。画像認識には対象の誤検出や見落としなどの問題が存在するため、必ずしも正しい情報が提供されるとは限らないが、それを差し引いても十分な精度向上が見込めるといえる。今後の課題として、より高度な課題に対するこの手法の適用可能性を検討することが挙げられる。今回は比較的単純な課題に対して事前知識を導入することの有意性を調査したが、複雑な課題に対しても適用できるかどうかについては議論の余地がある。また、事前知識の利用により期待される、未知の入力への対応力について検証することも今後の課題である。

参考文献

[1] Yannakakis, G. N. and Togelius, J.: *Artificial Intelligence and Games*, Springer (2018).

[2] Watkins, C. J. C. H.: Learning from delayed rewards, PhD Thesis (1989).

[3] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M. A.: Playing Atari with Deep Reinforcement Learning, *NIPS Deep Learning Workshop 2013* (2013).

[4] Winograd, T.: Understanding natural language, *Cognitive Psychology*, Vol. 3, No. 1, pp. 1 – 191 (online), DOI: https://doi.org/10.1016/0010-0285(72)90002-3 (1972).

[5] Branavan, S., Silver, D. and Barzilay, R.: Learning to Win by Reading Manuals in a Monte-Carlo Framework, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Association for Computational Linguistics, pp. 268–277 (2011).

[6] Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis, D. and Blunsom, P.: Grounded Language Learning in a Simulated 3D World, *arXiv:1706.06551* (2017).

[7] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105 (2012).

[8] Redmon, J., Divvala, S. K., Girshick, R. B. and Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection, *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 779–788 (2016).

[9] Redmon, J. and Farhadi, A.: YOLOv3: An Incremental Improvement, *CoRR*, Vol. abs/1804.02767 (2018).