強化学習を用いたゲームエージェントの評価

Evaluation of Game Agent Using Reinforcement Learning

リスク工学グループ演習１班

小清水亮太　宮澤一矢　原和希　HUANG YUMENG

アドバイザー教員　高安亮紀　遠藤靖典

01

研究背景

Background

02

研究目的

Purpose

03

研究手法

Method

04

研究結果

Result

05

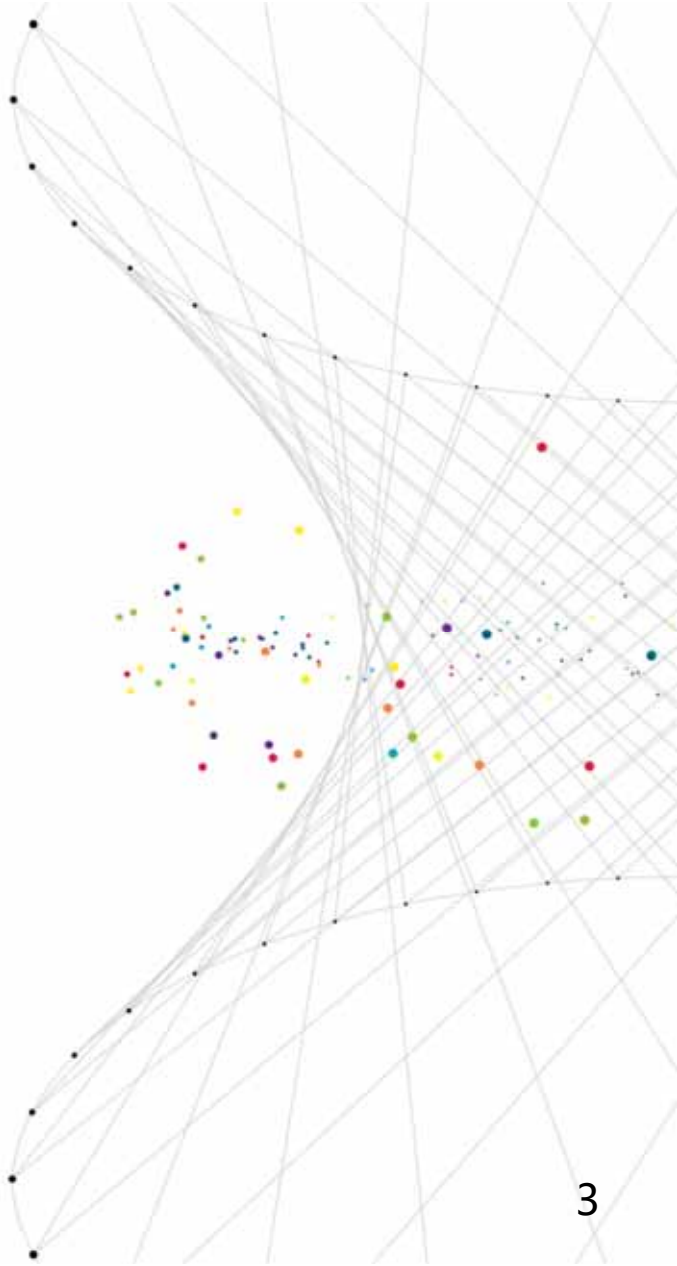
まとめ

Conclusion

01

研究背景

Background



第3次AIブームの到来
Arrival of the Third AI Boom

・近年AIの実用化が進み、第3次AIブームが到来[1].

・ビックデータのような大量のデータを用いることで人工知能（AI）自体が知識を取得する「機械学習」が実用化.

・知識を定義する要素を人工知能自身が自ら習得する「Deep Learning」が登場.

⇒第3次AIブームの到来により、機械学習を用いた研究が多くなされている

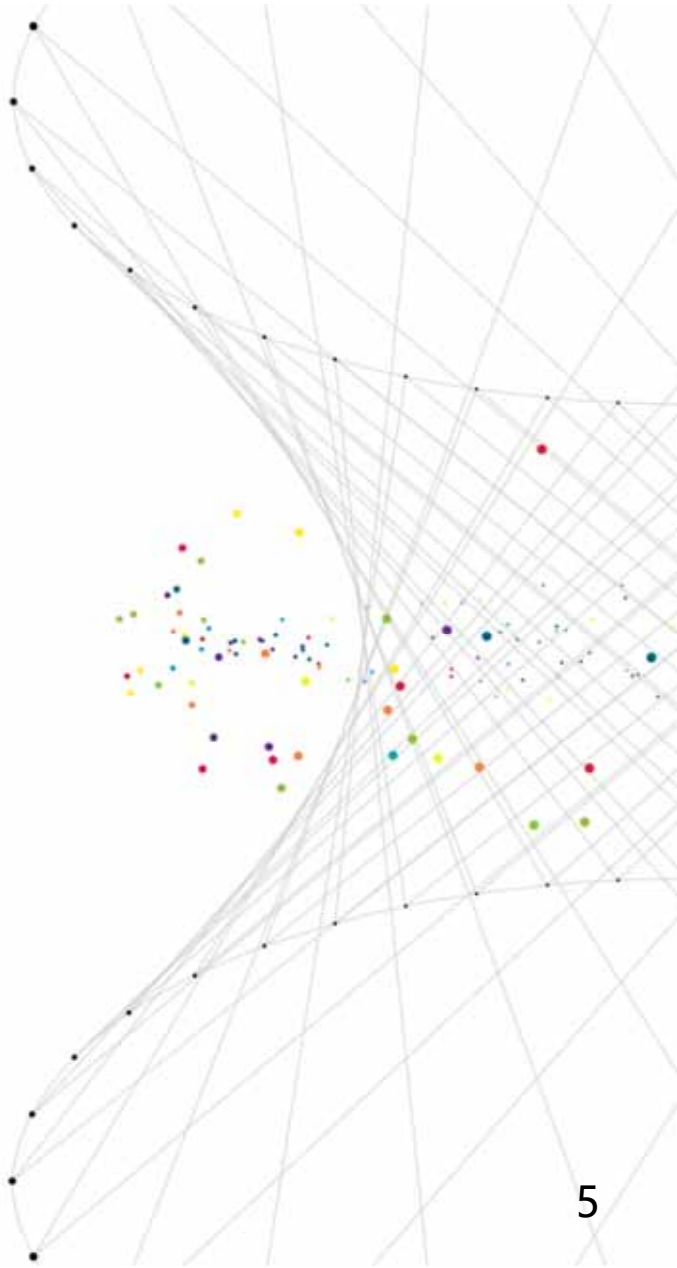
強化学習
Reinforcement Learning

・機械学習の手法の1つ.

・システム制御分野で多く利用.
ex) 自動運転技術、 自律ロボットの行動選択

・ゲーム開発の分野でも、動作の不具合の確認作業やゲームバランスの確認作業に利用.

・近年では、ゲームAIの作成にも用いられている.



ゲームAI
Game AI

・強化学習を用いることで強いゲームAIが提案されている.

・2015年、人間のプロ囲碁棋士に勝利する囲碁AIが誕生[2] .

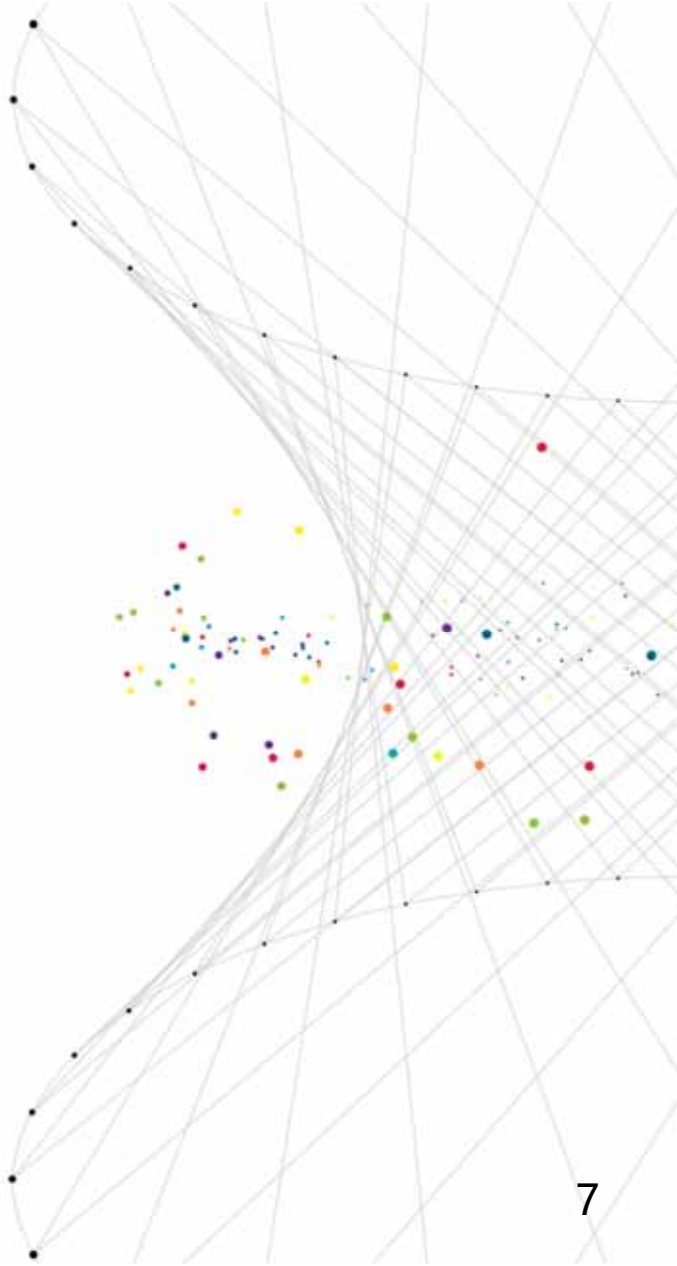
・2017年、人工知能PONANZAがプロ将棋棋士に勝利[3] .

プロの選手に勝利するゲームAIが誕生

02

研究目的

Purpose



・ゲームAIを作成するためには様々な強化学習手法が存在する.

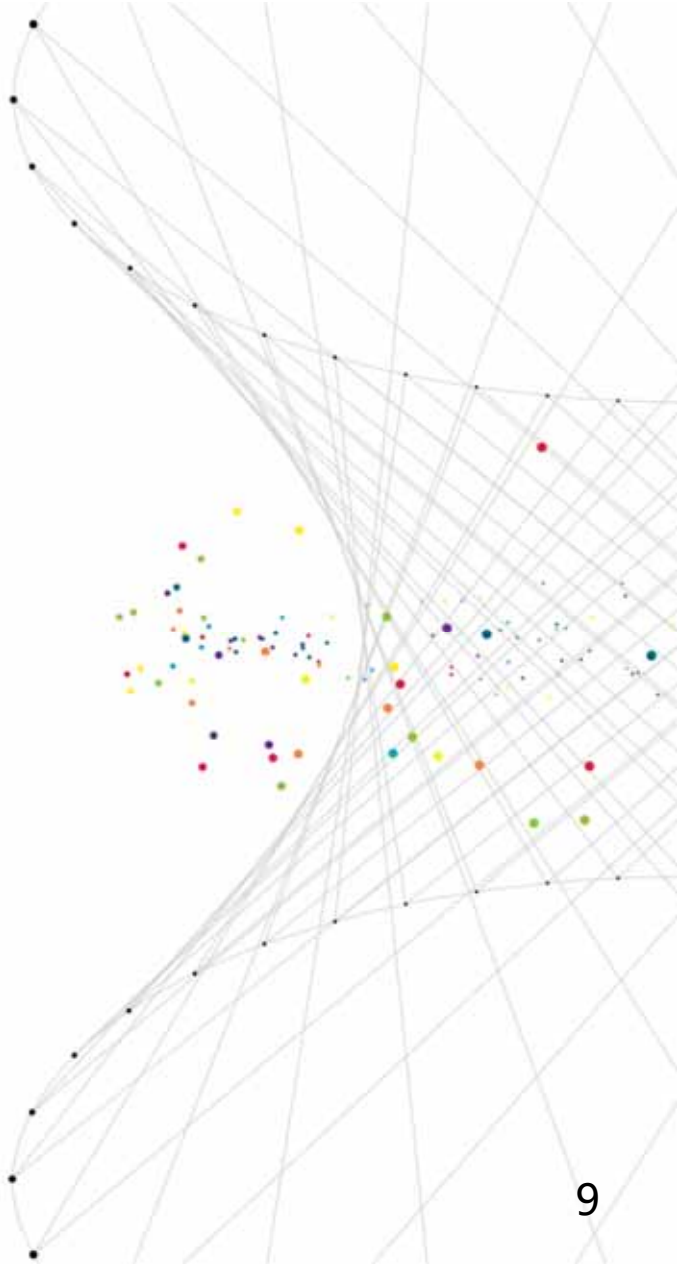
・強化学習手法間の結果や特徴の違いを把握することは、現実問題において適切な手法を選択するために重要.

本研究では、各種強化学習手法を用いて、ゲームAIを作成し、各種手法によるゲームAIの特徴を評価する.

03

研究手法

Method



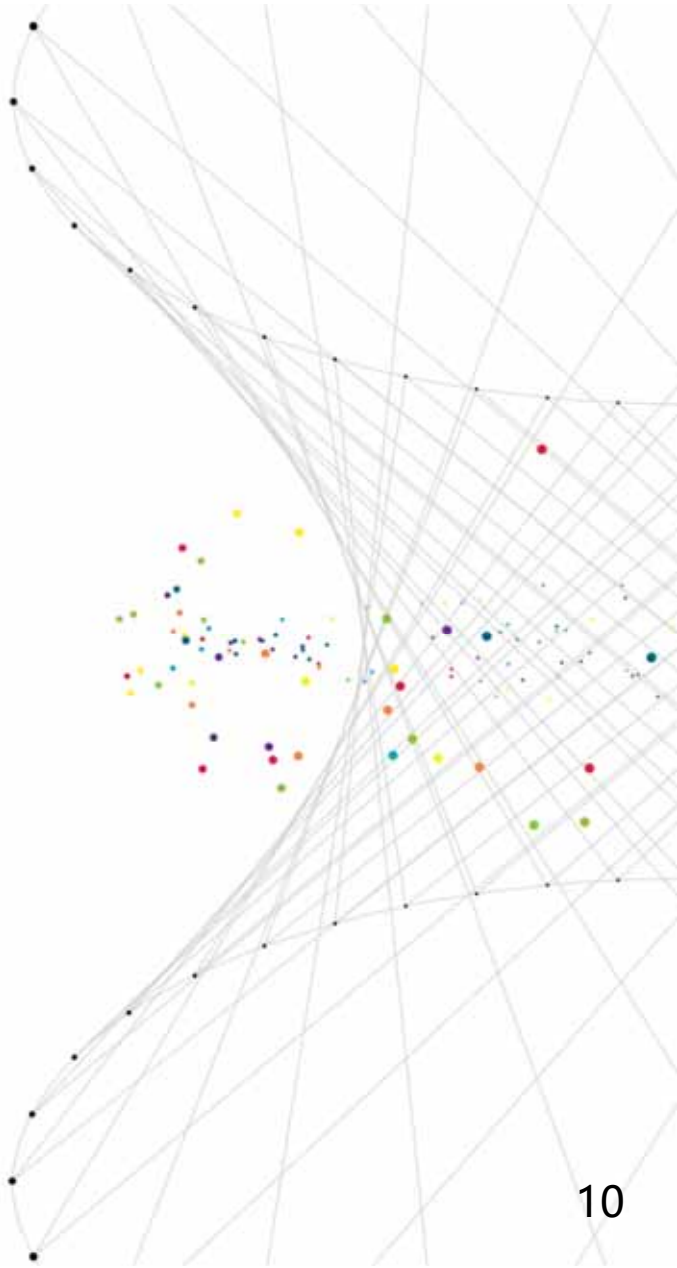
強化学習
Reinforcement Learning

・強化学習における重要な概念

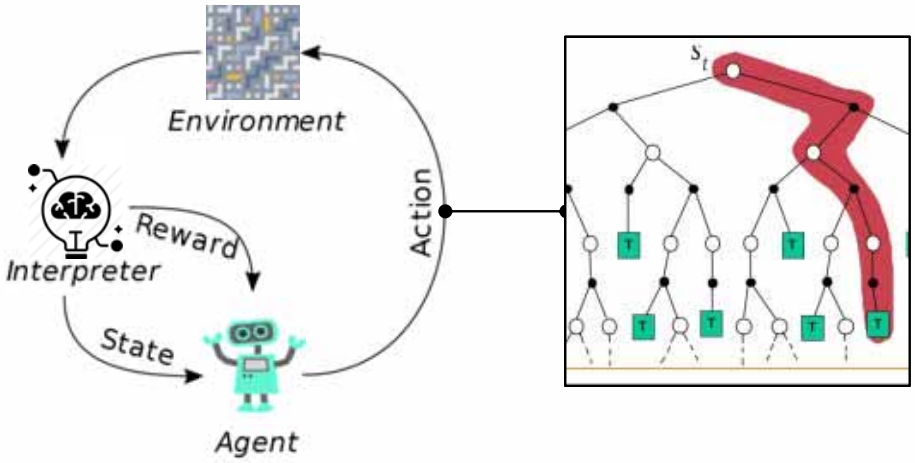
・状態 s :環境から取得される現在の状態
ex)エージェントの現在地、敵の位置

・行動 a :行動
ex)左右に移動する、ジャンプする

・報酬 R :状態sの時に行動aをした際に得られる利益



強化学習における流れ
Flow in reinforcement learning



図：強化学習のイメージ

Step1：環境から時刻tにおける状態s_tを観測

Step2：**方策**に従って行動a_tを決定

Step3：行動a_tによって状態s_{t+1}が決定

Step4：行動a_tによって報酬Rが決定

Step5：「Step1」に戻り、繰り返し

この方策を決めたい

行動価値関数
Action value function

$$Q = \begin{bmatrix} Q(S_1,a_1) & Q(S_1,a_2) & \cdots & Q(S_1,a_{|A|}) \\ Q(S_2,a_1) & Q(S_2,a_2) & \cdots & Q(S_2,a_{|A|}) \\ \vdots & \vdots & & \vdots \\ Q(S_t,a_1) & Q(S_t,a_2) & \cdots & Q(S_t,a_{|A|}) \\ \vdots & \vdots & & \vdots \\ Q(S_{|S|},a_1) & Q(S_{|S|},a_2) & \cdots & Q(S_{|S|},a_{|A|}) \end{bmatrix}$$

図：Q-tableの例

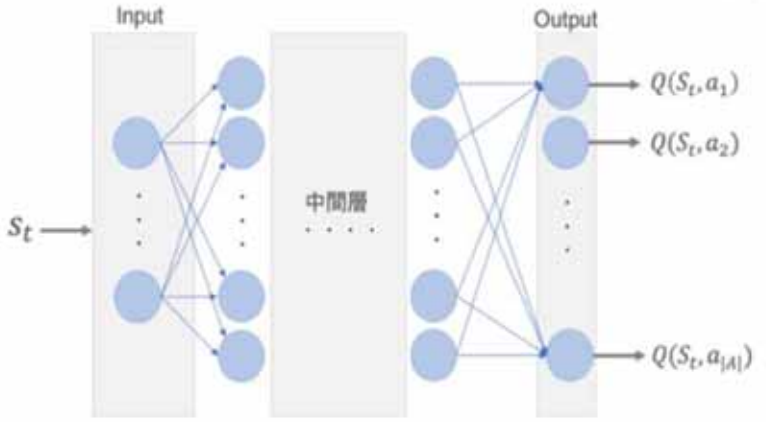
- 状態sでの行動aの価値を推定する関数
- 行動価値を得るための代表的手法→Q-learning
- 価値Q,状態s,行動a,報酬R,時間割引率γ,学習率αとする
- $Q(s,a) \leftarrow Q(s,a) + \alpha\,(R_{t+1} + \gamma max Q(s_{t+1},a_{t+1}) - Q(s,a))$
  に基づいて価値関数$Q(s,a)$を更新していく
- このQを用いてQ-tableを埋める
  →最適な方策の決定が可能に!

Deep Q Network(DQN) [4]

・状態の数が膨大→Q-tableで表現しきれない
・Deep Q Network(DQN)
…$Q(s,a)$をディープニューラルネットワークにより近似

$$Q = \begin{bmatrix} Q(S_1,a_1) & Q(S_1,a_2) & \cdots & Q(S_1,a_{|A|}) \\ Q(S_2,a_1) & Q(S_2,a_2) & \cdots & Q(S_2,a_{|A|}) \\ \vdots & \vdots & & \vdots \\ Q(S_t,a_1) & Q(S_t,a_2) & \cdots & Q(S_t,a_{|A|}) \\ \vdots & \vdots & & \vdots \\ Q(S_{|S|},a_1) & Q(S_{|S|},a_2) & \cdots & Q(S_{|S|},a_{|A|}) \end{bmatrix}$$

近似



図：DQNの例

Deep Q Network(DQN)

・最適な行動選択をする$Q_\theta(s,a)$,行動の結果s'でとるべき行動a'の価値を評価する$Q_\pi(s',a')$を用いる.

・DQNの更新式は以下のようになる

$$Q_\theta \leftarrow Q_\theta + \alpha(R(s,a) + \gamma max Q_\pi(s',a') - Q_\theta(s,a))$$

・明確な教師データが用意できないため、以下の式を教師データとして学習する

$$target_{DQN} = R(s,a) + \gamma \max(Q_\pi(s',a'))$$

Double DQN(DDQN) [5]

・DQNにおける$target_{DQN}$を改良した手法

・DQNでは行動選択と価値の評価で同じ関数$Q_\pi(s', a')$を使用
→ $Q_\pi(s', a')$に誤差があった際に過大評価

・行動選択と価値の評価を別の関数で行う

・これにより過大評価を抑制

Dueling Network[6]

・Q関数は，状態sのみできまる情報と行動aによって決まる情報に分離できる

・この点に注目し、これら2つ情報を分けて学習する手法

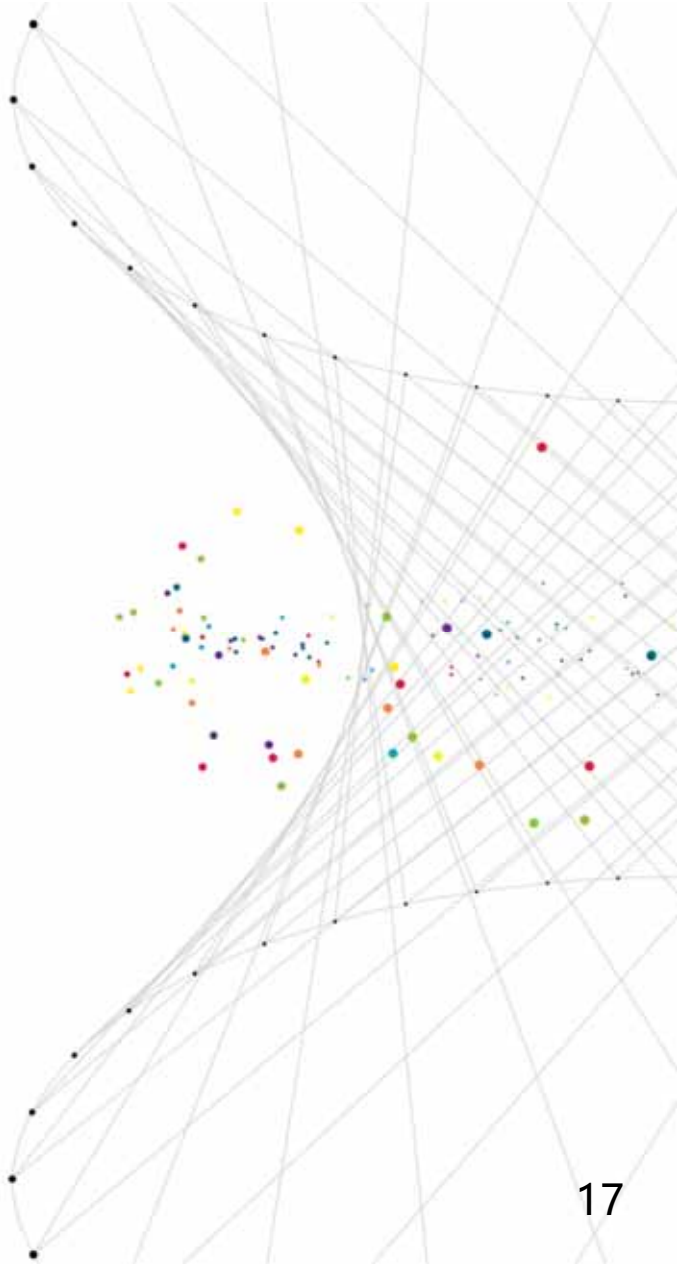
・DQN, DDQNと組み合わせて使用することが可能

・収束を早めたりパフォーマンスの向上が期待できる

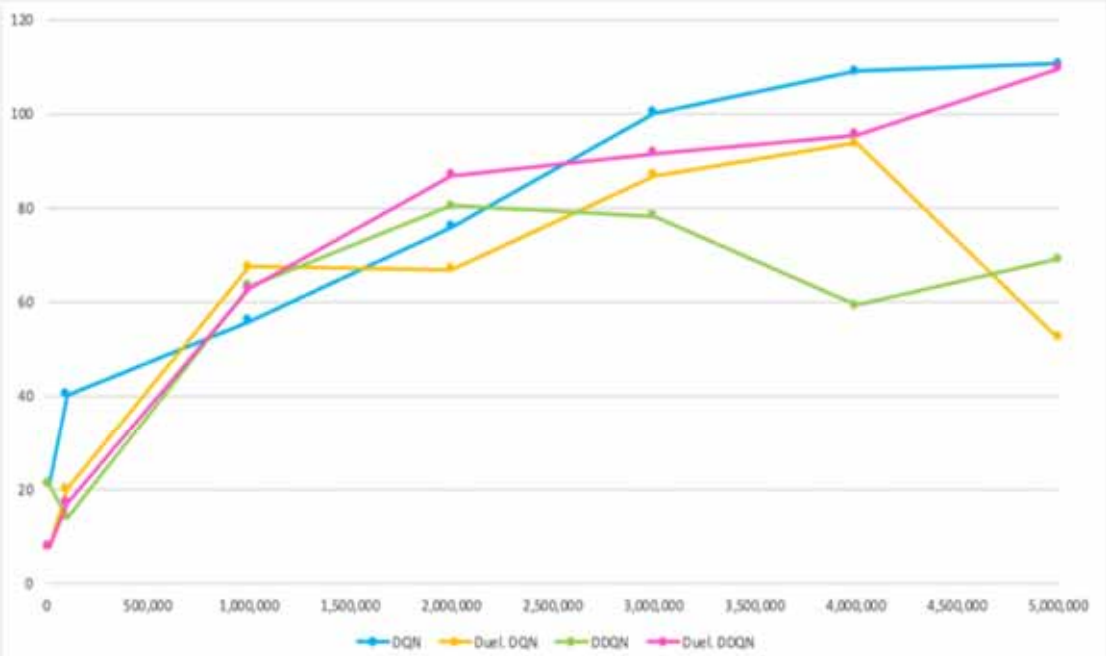
04

研究結果

Result



比較結果
Comparison result



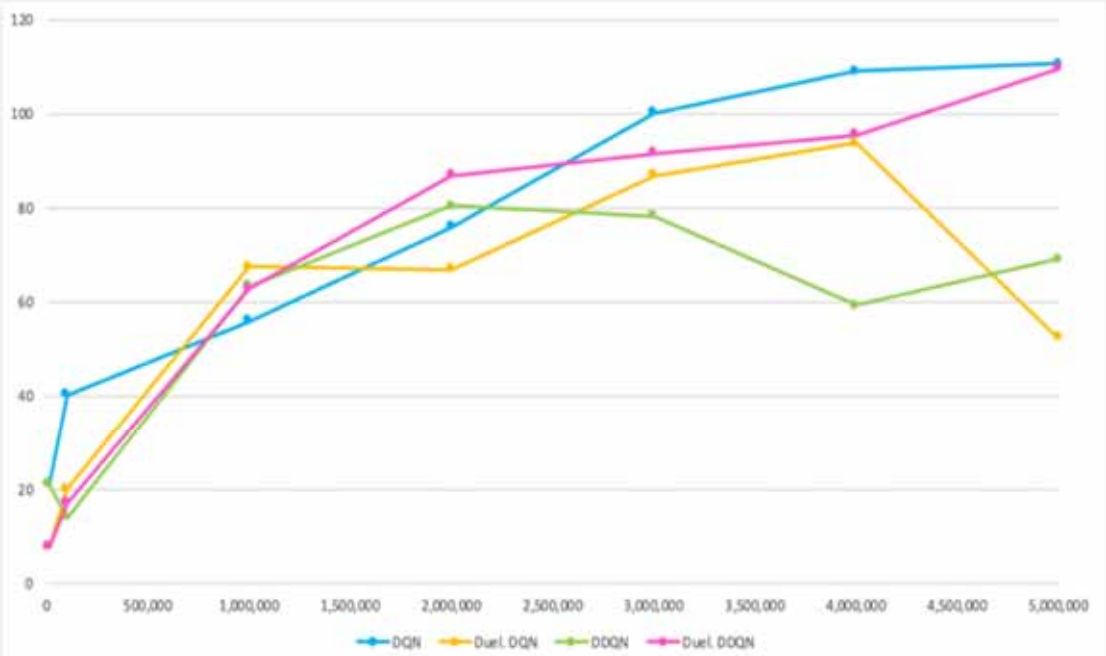
Step数	10,000	100,000	1,000,000	2,000,000	3,000,000	4,000,000	5,000,000
DQN	21	40.3	56	75.9	100.1	109	110.6
Duel. DQN	7.8	20.1	67.5	66.8	86.9	93.8	52.4
DDQN	21	14.1	63.3	80.4	78.2	59.2	69.1
Duel. DDQN	7.8	17.2	62.8	86.9	91.6	95.5	109.5

・強化学習の手法:
DQN　　DDQN
Duel. DQN　　Duel. DDQN

・Deep Learning実装
ライブラリである
KerasおよびKeras-RLを用いて実装

・一定のステップ数学習
(10,000～5,000,000回)

比較結果
Comparison result

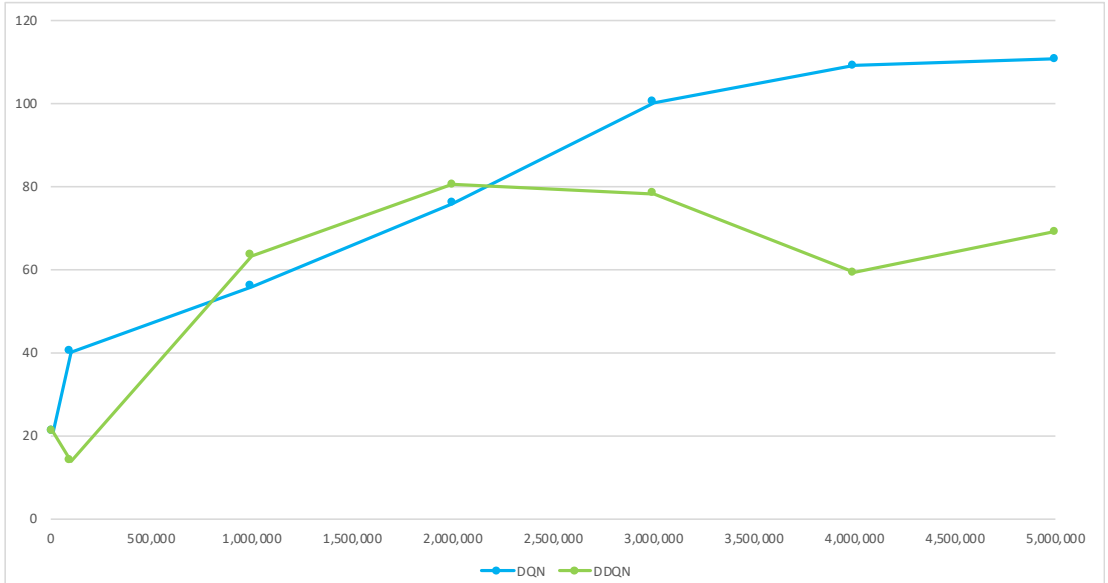


Step数	10,000	100,000	1,000,000	2,000,000	3,000,000	4,000,000	5,000,000
DQN	21	40.3	56	75.9	100.1	109	110.6
Duel. DQN	7.8	20.1	67.5	66.8	86.9	93.8	52.4
DDQN	21	14.1	63.3	80.4	78.2	59.2	69.1
Duel. DDQN	7.8	17.2	62.8	86.9	91.6	95.5	109.5

・学習済のエージェントで
10回テストした平均獲得報酬を
結果として記載

・横軸がステップ数
縦軸が獲得した報酬

DQN vs DDQN

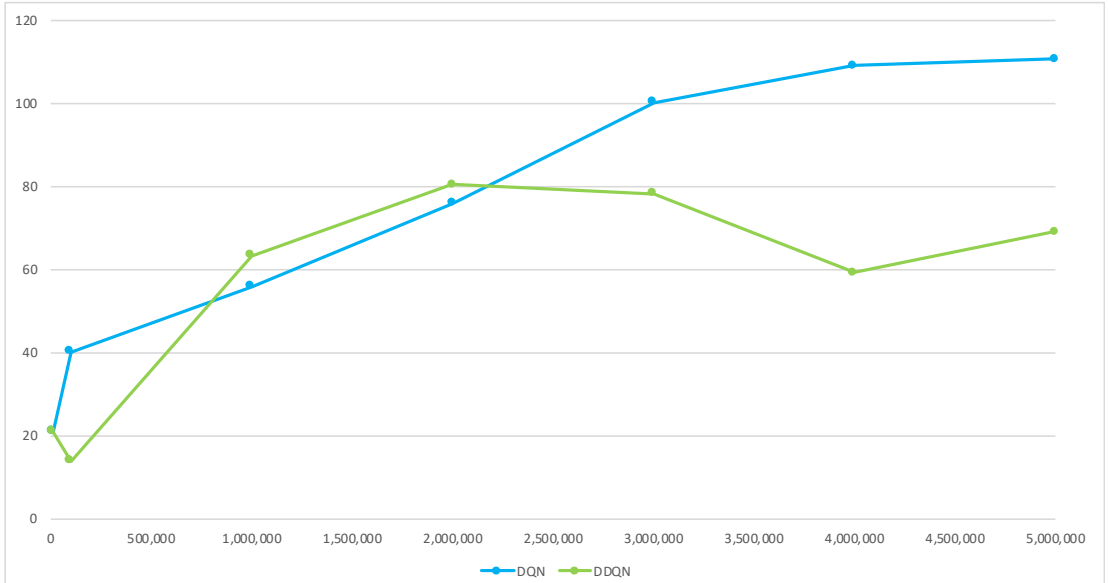


Step数	10,000	100,000	1,000,000	2,000,000	3,000,000	4,000,000	5,000,000
DQN	21	40.3	56	75.9	100.1	109	110.6
DDQN	21	14.1	63.3	80.4	78.2	59.2	69.1

- 強化学習の手法:
  DQN　　DDQN
- ほとんどのステップ数にて,
  DQNの方が報酬を多く獲得
- 既存研究:
  一部のゲームにおいては
  性能が悪化すると報告がある

我々が今回対象にしたMs. Pacmanも悪化する傾向にあることを確認した

DQN vs DDQN



Step数	10,000	100,000	1,000,000	2,000,000	3,000,000	4,000,000	5,000,000
DQN	21	40.3	56	75.9	100.1	109	110.6
DDQN	21	14.1	63.3	80.4	78.2	59.2	69.1

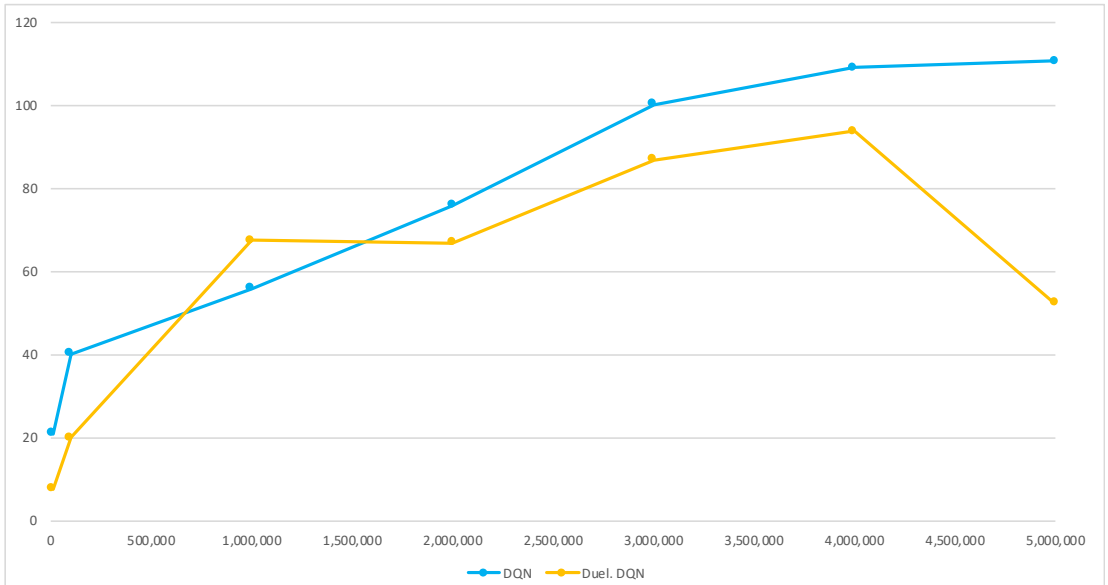
・強化学習の手法:
DQN　　DDQN

・ほとんどのステップ数にて,
DQNの方が報酬を多く獲得

・既存研究:
一部のゲームにおいては
性能が悪化すると報告がある

※ DDQNにおいてニューラルネットワークの更新間隔を調整することによるパフォーマンス向上も報告
→ 今後はDDQNを用いてゲームごとに適切なパラメータ探索を行う必要がある

DQN vs Duel. DQN



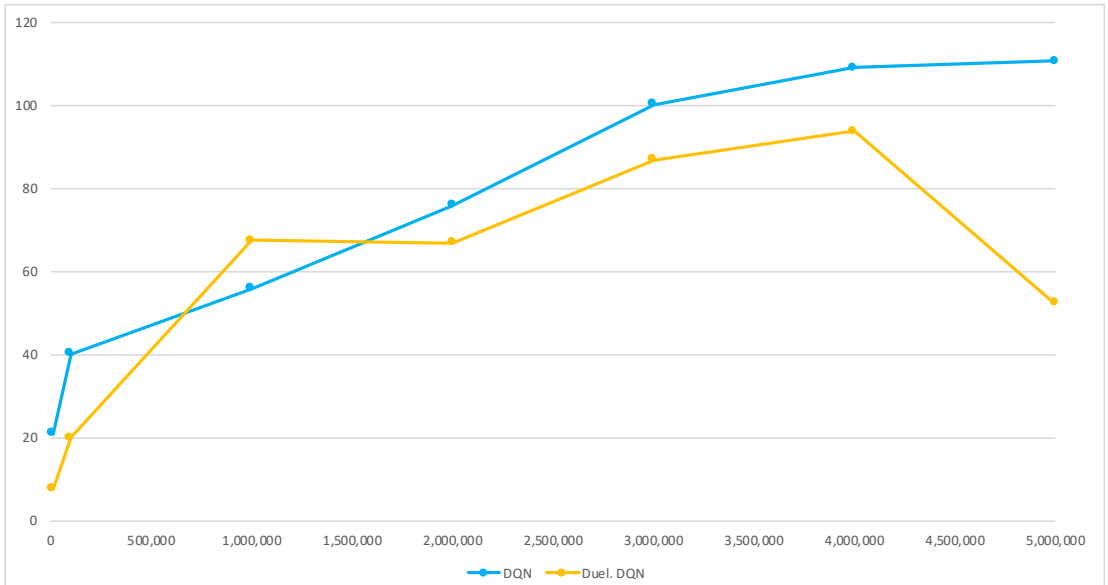
Step数	10,000	100,000	1,000,000	2,000,000	3,000,000	4,000,000	5,000,000
DQN	21	40.3	56	75.9	100.1	109	110.6
Duel. DQN	7.8	20.1	67.5	66.8	86.9	93.8	52.4

・強化学習の手法:
DQN　　Duel. DQN

・比較結果:
- DQN　　→学習が安定している
- Duel. DQN → 学習が不安定

-ハイパーパラメータの数が多く，本検討ではハイパーパラメータの調整が不足している為

DQN vs Duel. DQN



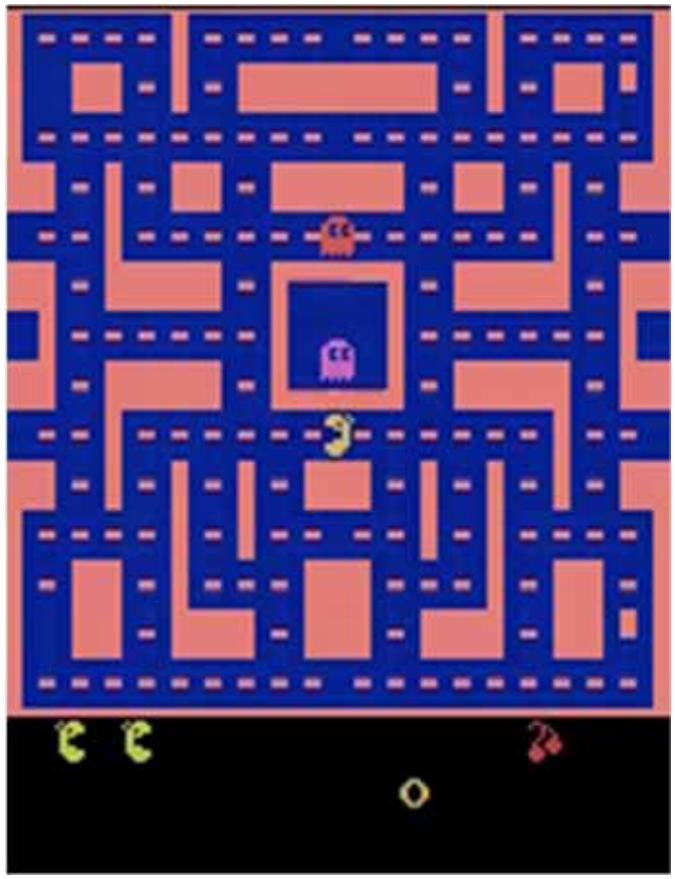
Step数	10,000	100,000	1,000,000	2,000,000	3,000,000	4,000,000	5,000,000
DQN	21	40.3	56	75.9	100.1	109	110.6
Duel. DQN	7.8	20.1	67.5	66.8	86.9	93.8	52.4

- Duel. DQN
  - 4,000,000ステップで最高報酬
  - それ以降は過学習によって，報酬が低下

→ **収束を早める効果**を確認

Dueling Networkの使用：収束を早めることが出来るが，ハイパーパラメータの探索がより重要になる

作成エージェントによるプレイ動画
Video



・入力画像： RGB（カラー画像として使用)

・使用手法: DQN

・学習ステップ数: 500万

・カラー画像の使用によって，パワークッキーを使った敵の撃退が増加した

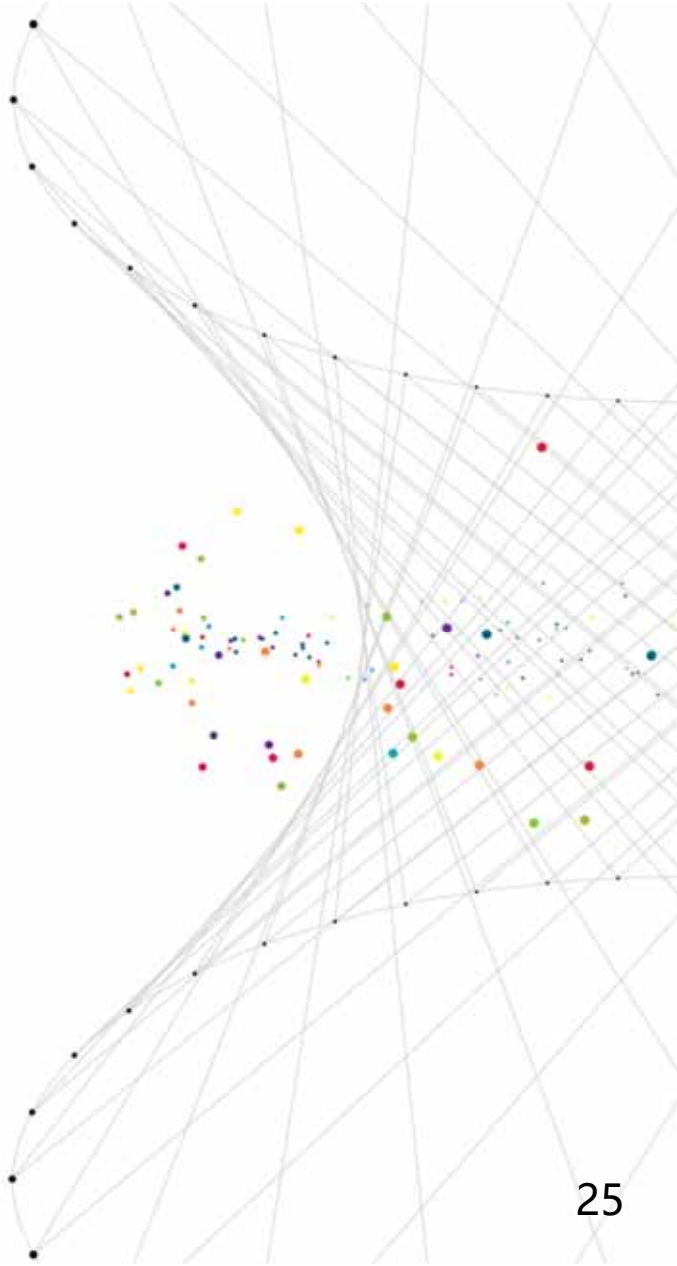
・4280点獲得（ 中間発表時では最高1370点)

・学習時間の増加がデメリット

05

まとめ

Conclusion

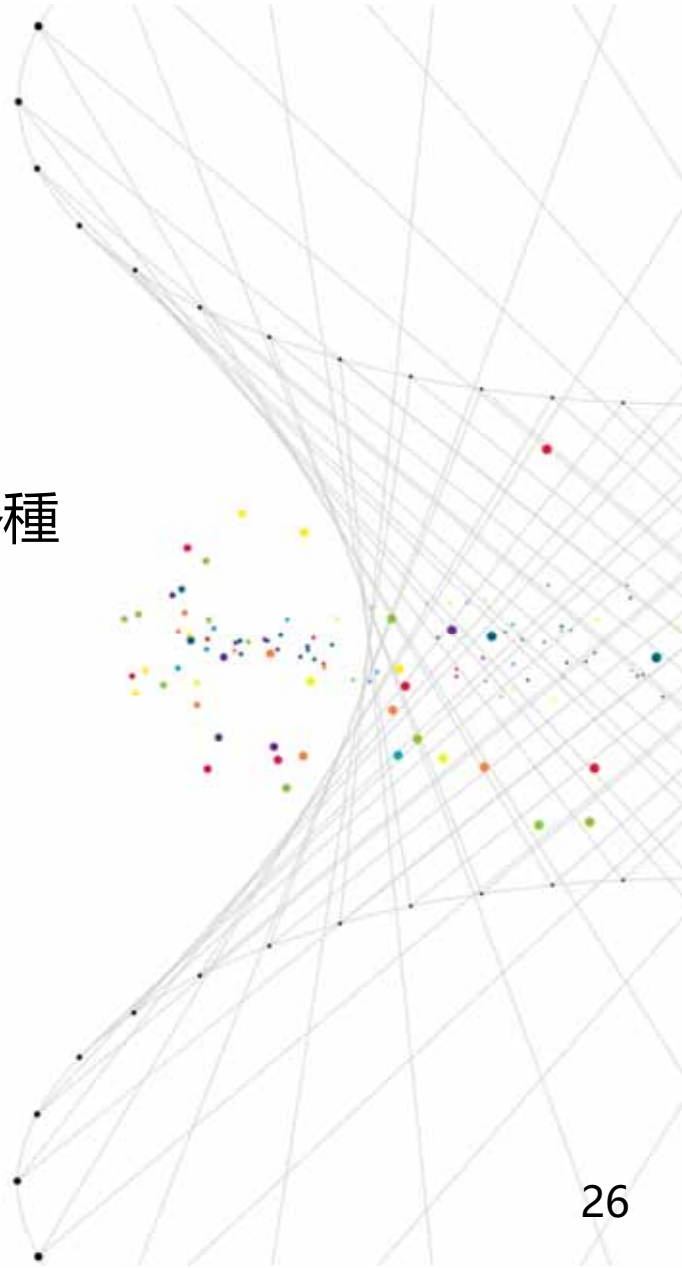


まとめ
Conclusion

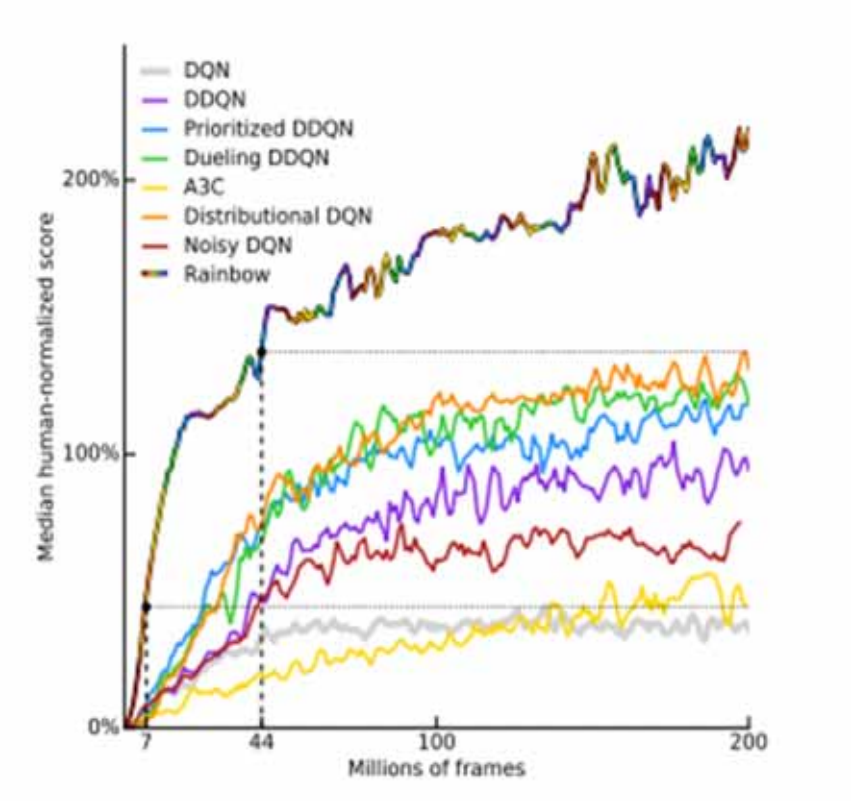
・本研究では、 Ms. Pac-Manというゲームを題材に各種強化学習手法を用いて実験を行った.

・DQNを用いて5,000,000ステップ学習させたものが平均獲得報酬が最も大きいという結果になった.

・今回の問題における各手法の結果や特徴について評価することが出来た.



今後の課題
Future Work



図：Rainbow

・今回実験をしていない手法による評価
ex)　Rainbow [7]

・ハイパーパラメータや層の
適切なチューニングの検討

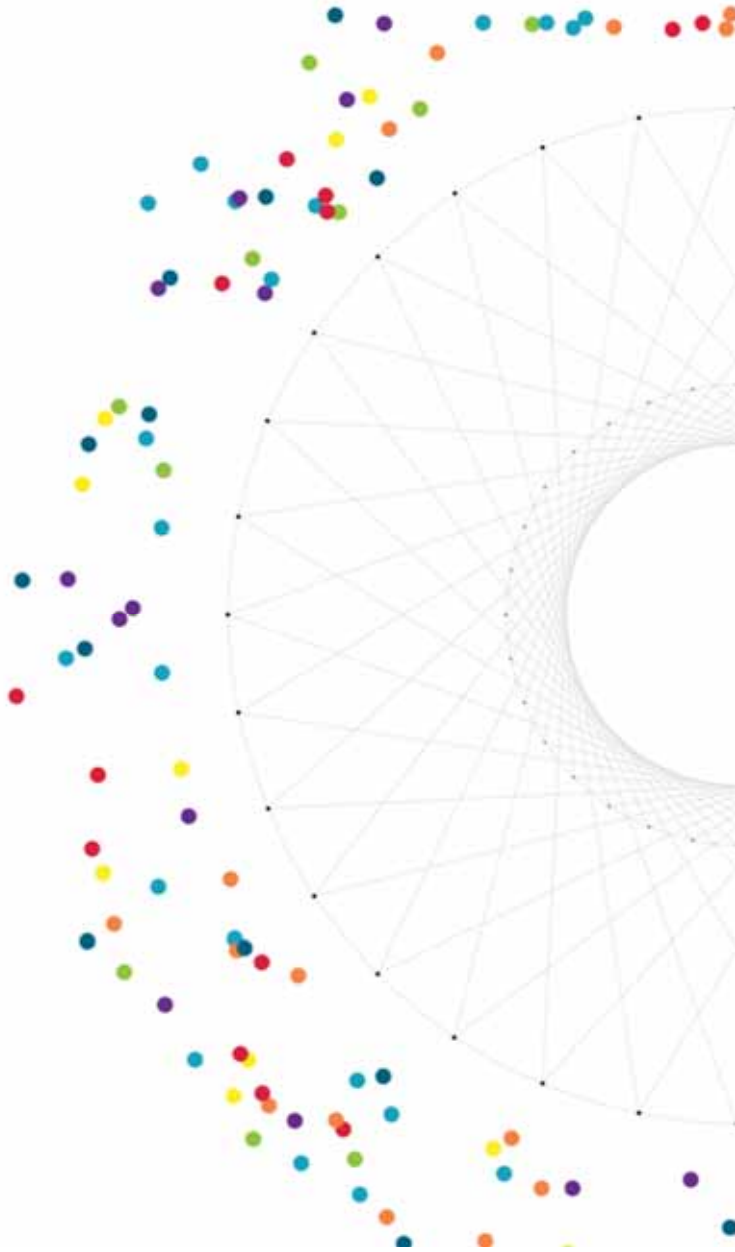
参考文献 1
References

[1]総務省. 平成28年版 情報通信白書｜人工知能（AI）の研究の歴史, http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc142120.html .

[2] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”, Nature529, 484-489(2016).

[3] HEROZ株式会社, Ponanzaにおける強化学習とディープラーニングの応用, https://www.slideshare.net/HEROZ-JAPAN/ponanza-83900718.

[4]Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540 (2015)

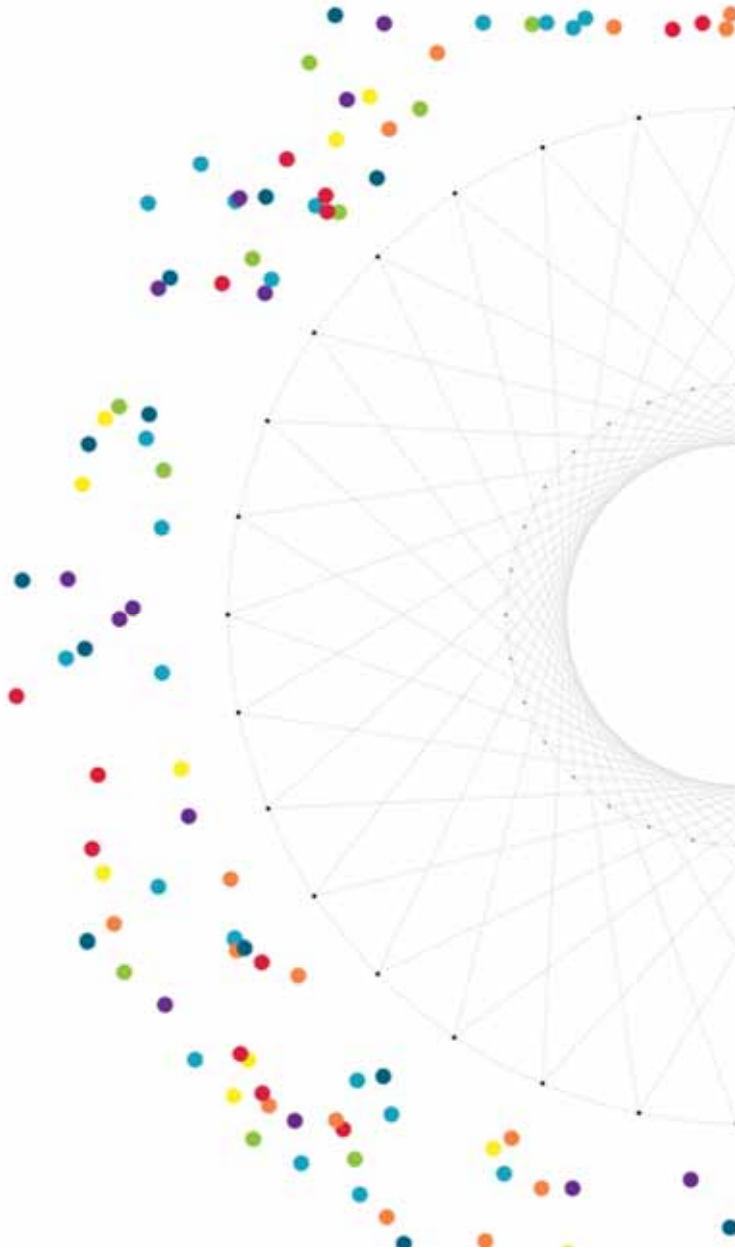


参考文献 2
References

[5] Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep reinforcement learning with double q-learning." Thirtieth AAAI conference on artificial intelligence. 2016

[6] Wang, Ziyu, et al. "Dueling network architectures for deep reinforcement learning." arXiv preprint arXiv:1511.06581 (2015).

[7]Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, David Silver. “Rainbow: Combining Improvements in Deep Reinforcement Learning ” arXiv:1710.02298v1 [cs.AI] 6 Oct 2017



ご清聴ありがとうございました

Thank you for watching