二次最適化の強化学習への適用

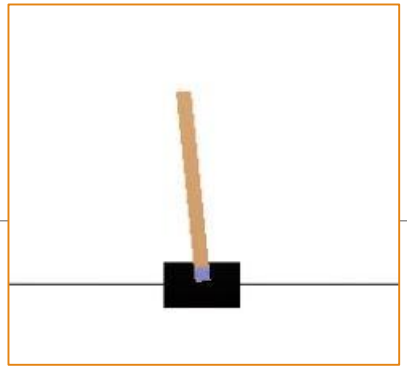
構想発表会

横田理央研究室

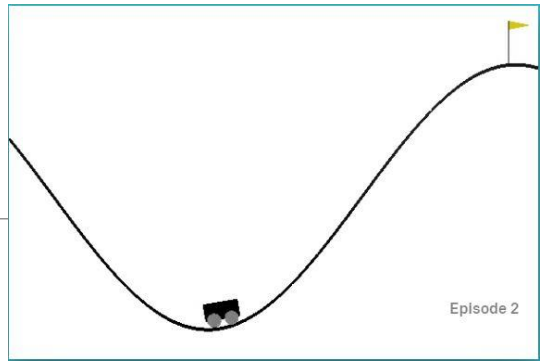
18M30574　桑村祐二

概要

CartPole-v1



MountainCar-v0



- 導入・問題提起
- マルコフ決定過程(MDP)
- 強化学習テクニック (Advantage Function, Actor-Critic)
- ACKTR新規性 (K-FAC)
- 実験結果・課題
- 直近の研究内容
- スケジュール

ACKTR : Actor Critic using Kronecker-Factored Trust Region

Yuhuai Wu et al. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. 2017

導入・問題提起

・DNNを用いた深層強化学習は空間的かつ複雑な制御に貢献

・タスクの難化・連続化により学習時間の増大

・最適化手法として確率的勾配降下法(SGD)が主流だが非効率

・単純な分散では分散数に見合わない程度の高速化 (Mnih et al, ICML2016)

・既存手法 TRPO (Schulman et al, ICML2015) は多数のサンプルを必要とする

→ 大規模モデルには不適

・K-FACならbatch sizeに関係なく適切に近似可能かつSGDとコスト同等

マルコフ決定過程(MDP)

・強化学習の場合、これに割引率が加わる　$\langle S, A, P, r \rangle + \gamma$(割引率)

S : 状態集合 , A : 行動集合 , P : 遷移確率関数 , r : 報酬

①現在の状態 $s \in S$, 方策関数 $\pi(a|s;\theta)$ を元に行動 $a \in A$ を選択

②Agent は Environment から報酬 $r(s,a)$ を得る

③遷移確率関数 $P(s'|s,a)$ より次の状態 s' へ遷移する

割引率を考慮した将来報酬和(割引報酬和)を最大化する行動を選択できるか？

$$J(\theta) = \sum_{i=0}^{\infty} \gamma^i r(s_i, a_i) \quad \big(\approx V(s)\big) \ \big(\approx Q(s,a)\big)$$

強化学習テクニック

・状態によって行動の重要度が違うことを学びたい

状態価値関数 $V(s)$, 行動価値関数 $Q(s,a)$

両方使えば割引報酬和をより適切に表現できる

→ Advantage Function (Wang et al, ICML2016)

$$A(s,a) = Q(s,a) - V(s)$$

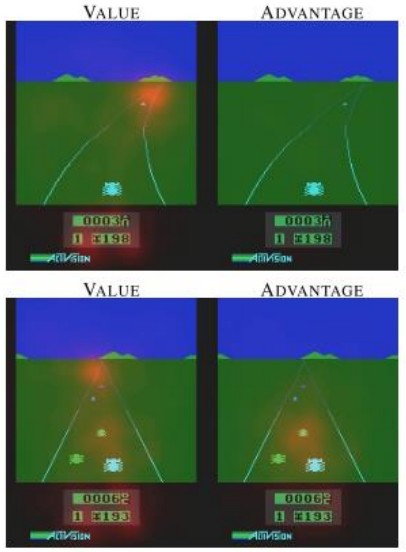


Figure 2. See, attend and drive: Value and advantage saliency maps (red-tinted overlay) on the Atari game Enduro, for a trained dueling architecture. The value stream learns to pay attention to the road. The advantage stream learns to pay attention only when there are cars immediately in front, so as to avoid collisions.

・得られる報酬の差が大きいと学習が安定しない

「行動決定(Actor)」と「状態行動評価(Critic)」を別々に学習

→ Actor-Critic (Sutton & Barto, MIT Press 1998)

$$\nabla_\theta J(\theta,\phi) = \mathbb{E}[\nabla_\theta log\pi(a|s;\theta)Q(s,a;\phi)]$$

ACKTR新規性

・SGDを用いた方策勾配法(前スライドは目的関数の工夫)

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla_\theta J(\theta, \phi)$$

・ブロック対角近似を用いてFisher情報行列を近似したK-FAC

(Martens and Grosse, ICML2015)

・これを更新式として用いたのがACKTR

$$\left[F(\theta^{(t)})\right]_{i,j} = \mathrm{E}[a_{i-1}a_{j-1}^T \otimes g_i g_j^T] \approx E[a_{i-1}a_{j-1}^T] \otimes E[g_i g_j^T]$$

$$\left[\hat{F}(\theta^{(t)})\right]_{i,j} := E[a_{i-1}a_{j-1}^T] \otimes E[g_i g_j^T]$$

$$\theta^{(t+1)} = \theta^{(t)} - \hat{F}(\theta^{(t)})^{-1} \nabla_\theta J(\theta, \phi)$$

実験結果・課題

①Arcade Learning Environment

畳み込み層：3

全結合層　：1

6種類のタスク全てReward合計

ACKTR > A2C > TRPO

②MuJoCo (共に”Platform”)

畳み込み層：2

全結合層　：1

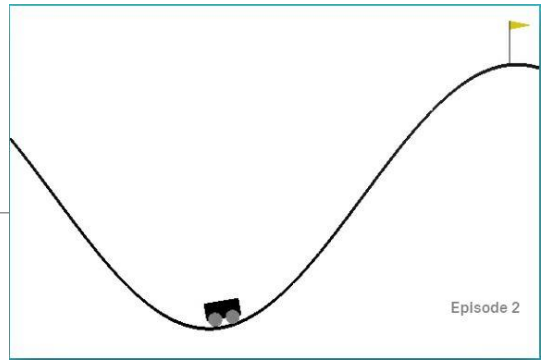
タスクに依りトップが異なる

(8タスク中ACKTRトップは2タスク)

・(画像認識等と比べて)モデルが小さすぎる　→　モデルの再検討
・ACKTR以外のモデルが一部異なる　→　モデルの統一
・Episode数の異なるReward比較　→　Episode数の統一

・比較対象アルゴリズムが少ない(3つ)　→　高性能アルゴリズムとの比較
・Parameterizeタスクで映えないACKTR　→　K-FAC適用の改善

直近の研究内容 (1/2)



・「MountainCar-v0」(OpenAI Gym, Classic Control)

・車を(旗のある)山頂へと移動させるのが目標

$S = \{position, velocity\}$

$A = \{left, idle, right\}$

$r = \{0: arrived,\ -1: else\}$

・200回以内のtimestepで報酬和を算出

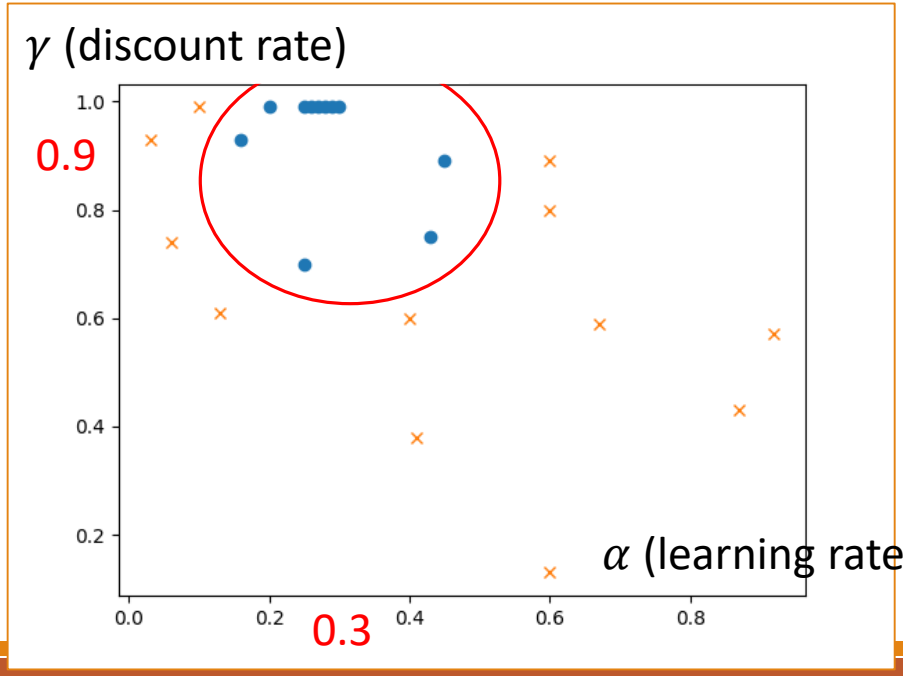
・Q-learning (α: learning rate)

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left\{r(s,a) + \gamma \max_{a' \in A} Q(s',a') - Q(s,a)\right\}$$

直近の研究内容 (2/2)

- 強化学習は結果の変動が大きい (Henderson et al, AAAI2018)
  → Hyper Parameter を決め打ちしない(唯一としない)
- 「Episode10区分最低1つ到達率90%」という閾値で良悪を分類
- $HyperParameter = \{\alpha, \gamma\}$

- 右図のように境界が引けそう…
- $(\alpha, \gamma) = (0.3, 0.9)$ で確認
  → 学習が破綻せずに精度良
  (前スライドのグラフ)



スケジュール

時期	主な研究(学業)内容	補足
2018年4月～9月	専門科目履修	文系科目は分散的に
10月～11月	Parameter Tuning 追試	データや手法を変えて比較
12月～2019年1月	強化学習の論文を収集	各論文を単語でラベリング
2月～3月	Q学習系での実験	classic control(OpenAI Gym)
4月～6月	方策型アルゴリズム整理	環境構築、実装済試験
7月～9月	アルゴリズム開発	K-FAC利用を重点的に
10月～12月	実験	データセットはMuJoCo想定
1月	修論仕上げつつ追試	

※論文収集と修論執筆は随時行う。

(補足スライド)
自然勾配近似法の強化学習への応用

構想発表会

横田理央研究室

18M30574　桑村祐二

概要

- 本発表のまとめと論文詳細
- Advantage Function / Actor-Critic / 自然勾配法 / K-FAC
- 機械学習の種類
- 教師あり学習 / 教師なし学習 / 強化学習
- 頻出用語の補足
- Q学習から方策型にシフトした理由
- OpenAI Gym / Montezma’s Revenge / その他研究タスク
- 予定とする研究内容
- 当研究による展望

本発表のまとめと論文詳細

・自然勾配を近似して用いる二次最適化手法K-FAC

(James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. 2015)

・強化学習アルゴリズムACKTRに用いられる

(Yuhuai Wu et al. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. 2017)

・しかしParameterベースのタスクでは他に劣る性能

(Peter Henderson et al. Deep Reinforcement Learning that Matters. 2017)

(Scott Fujimoto et al. Addressing Function Approximation Error in Actor-Critic Methods. 2018)

Advantage Function (集約前1)

・割引報酬和を定義する価値関数は2種類ある

状態価値関数 $V(s)$, 行動価値関数 $Q(s,a)$

Vのみ　→　行動で報酬が大きく変わる場合(直立時など)

Qのみ　→　状態で報酬が決まる場合(倒れる直前など)

両方使えば割引報酬和をより適切に表現できる

→　Advantage Function (係数工夫するなどの亜種有)

$$A(s,a) = Q(s,a) - V(s)$$

Actor-Critic (集約前1)

・非Actor-Criticな方策勾配法における割引報酬和（損失関数）の微分

$$\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta log\pi(a|s;\theta)Q(s,a;\theta)]$$

・得られる報酬の差が大きいと学習が安定しない

・対してActor-Criticではモデルを分けることで問題を解決

・「行動決定(Actor)」と「価値関数(Critic)」を別々に学習

$$\nabla_\theta J(\theta,\phi) = \mathbb{E}[\nabla_\theta log\pi(a|s;\theta)Q(s,a;\phi)]$$

(Sutton R. and Barto A. Reinforcement Learning: an Introduction. 1998)

自然勾配法 (集約前2)(acktr更新式追加)

・最適化手法の一種

(S. I. Amari. Natural gradient works efficiently in learning. Neural Computation, 10(2):251–276, 1998.)

・従来の手法ではプラトーや局所的最適解といった問題が発生
- →自然勾配法を深層学習に適用

$$\theta^{(t+1)} = \theta^{(t)} - G(\theta)^{-1}\nabla_{\theta}J(\theta)$$

$G(\theta)$: パラメータ θ に適した空間を定義した行列

・損失関数がユークリッド空間上では適切に表現できないと仮定

・既存手法を行列Gによって一般化
- →行列Gを定義する必要がある

K-FAC (1/2) (集約前2)

- 自然勾配法の近似手法、クロネッカー因子分解を用いる
  - (James Martens, Roger Grosse :
  - Kronecker-factored Approximate Curvature : 2015)
- 行列$G(\theta)$にフィッシャー情報行列$F(\theta)$を用いる (二次微分を用いる)

$$F(\theta) = \mathbb{E}\left[\nabla_\theta log\pi(a|s;\theta)\left(\nabla_\theta log\pi(a|s;\theta)\right)^T\right]$$

- パラメータの個数を N とすると、行列$F(\theta)$のサイズは $N \times N$
  - →精確な逆行列計算は現実的でない
  - (GoogLeNet : N≈6.8 × 10^6 , AlexNet : N≈6.2 × 10^7 , VGG-16 : N≈1.4 × 10^8)
  - →逆行列をブロック対角で近似して計算

K-FAC (2/2) (割愛予定)

a_i : (順伝播時における) i番目の層の入力

g_i : (逆伝播時における) i番目の層に伝播する微分値

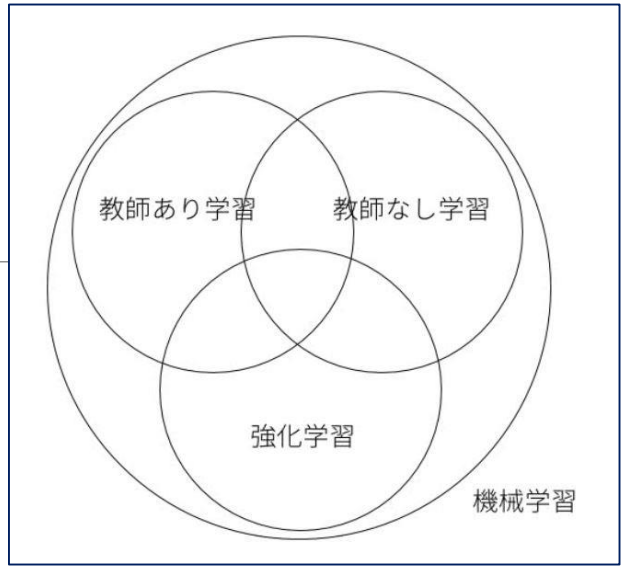
$\left[F(\theta^{(t)})\right]_{i,j}$: 行列 $F(\theta^{(t)})$ の i,j 成分

$$\left[F(\theta^{(t)})\right]_{i,j} = \mathrm{E}[a_{i-1}a_{j-1}^T \otimes g_i g_j^T] \approx E[a_{i-1}a_{j-1}^T] \otimes E[g_i g_j^T]$$

$$\left[\hat{F}(\theta^{(t)})\right]_{i,j} := E[a_{i-1}a_{j-1}^T] \otimes E[g_i g_j^T]$$

$$\theta^{(t+1)} = \theta^{(t)} - \hat{F}(\theta^{(t)})^{-1}\nabla_\theta J(\theta)$$

機械学習の種類



・教師あり学習 (Supervised Learning)

入力と出力の関係を学習、「分類」「予測」

・教師なし学習 (Unsupervised Learning)

データの構造を学習、「クラスタリング」「次元削減」

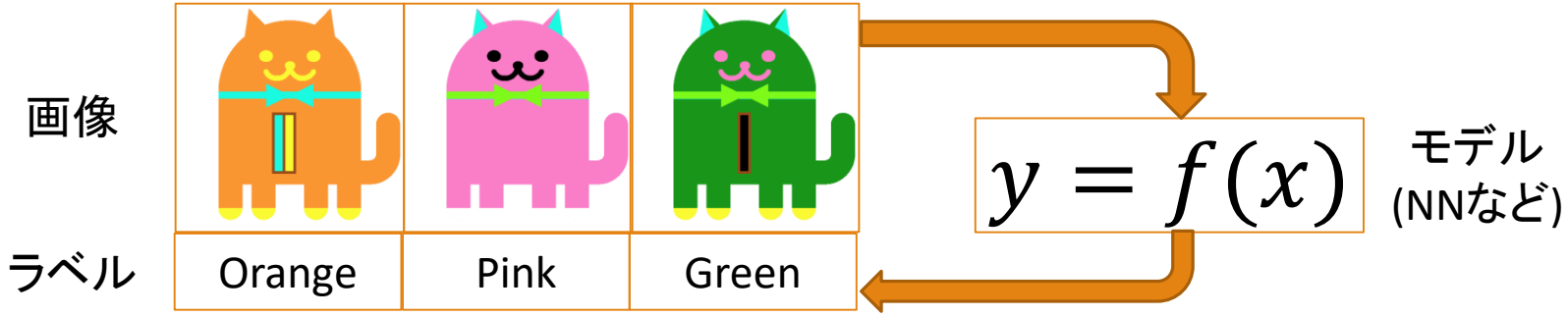
・強化学習 (Reinforcement Learning)

「状態」から「行動」を選択し「報酬」を受け取る、「報酬」の最大化

確率モデル：マルコフ決定過程(MDP) {状態、行動、遷移、報酬}

教師あり学習について

・出力ラベルを付与したデータセットを対象とすることが多い



・特定の属性を出力とするモデル構築も教師あり学習に含まれる

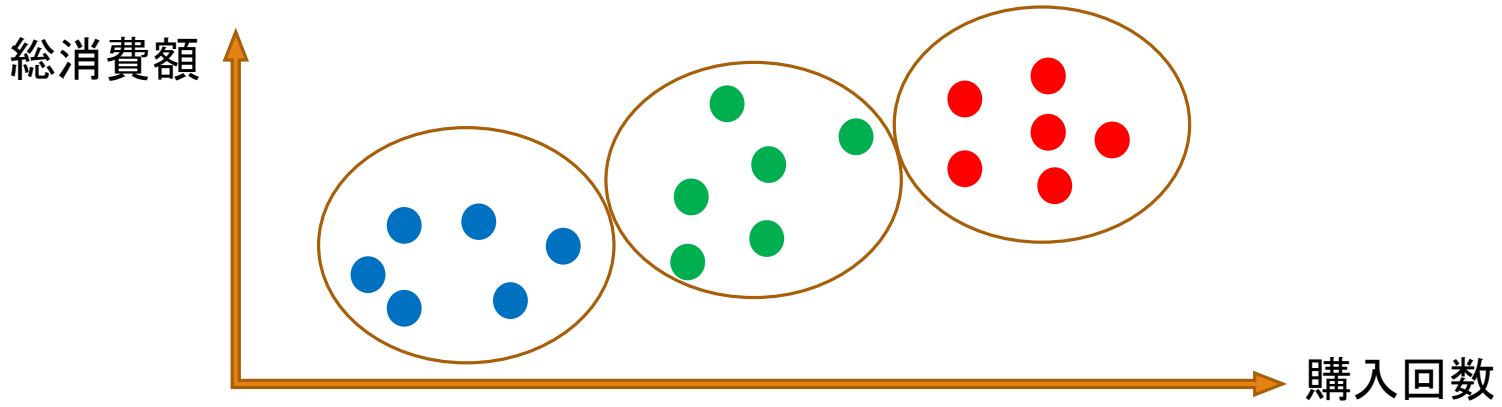
入力

客数(人)	価格(円)	気温(℃)	利益(円)
10,000	1,890	12	100,000
20,000	1,920	18	300,000
40,000	2,000	14	250,000
⋮	⋮	⋮	⋮

出力

教師なし学習について

・データの「構造」を元にカテゴライズを行う「クラスタリング」



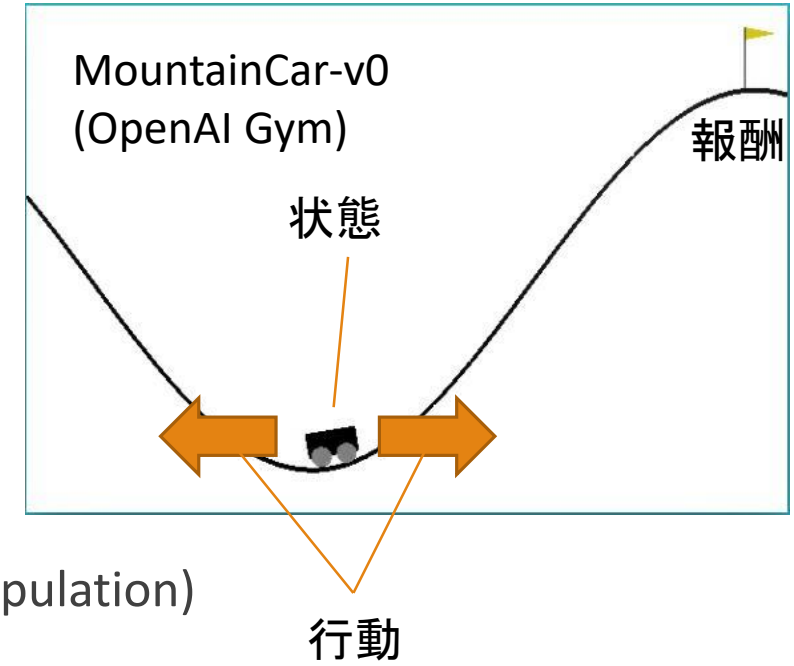
・データが大量な場合に用いられることが多い

・多次元なデータの場合、重要度の低い要素を除く「次元削減」

強化学習について

・「状態」から「行動」を選択し「報酬」を受け取るマルコフ決定過程(MDP)

報酬の合計が最大となるように学習



・ゲームのイメージが強い？

(最終スコアを最大化)

・報酬を適切に設計すれば幅広く応用

例1))指先の動きを学習(Dexterous Manipulation)

例2))資産運用等のファイナンス

頻出用語の補足

・Q関数(行動価値関数)　Q(s,a)

「状態」と「行動」を引数にとり将来得られる報酬和を示す

(精確に求めるのは現実的でないので近似するのが一般的)

・方策関数　π(a|s), π(s)

「状態」から「行動」を選択する確率分布

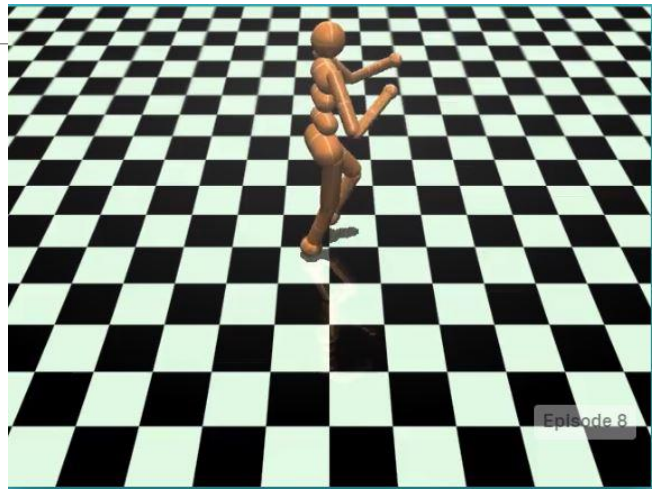
(「状態」を引数として「行動」を出力する関数を指す場合もある)

(Parameterベースでは前者の方が一般的)

(共有時)(Q関数は方策関数に依存するので「Q^{π}」と表すことがある)

Q学習から方策型にシフトした理由

Humanoid-v2
(MuJoCo, OpenAI Gym)



・簡単なタスクではQ学習系で間に合ってた

(行動が有限 or 有限にしても精度良い場合)

(Q関数を最大とする行動選択が容易)

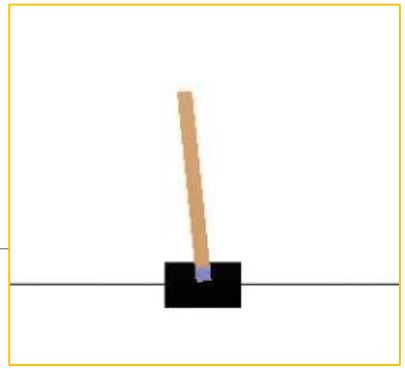
・実社会への応用に問題あり

行動がParameterベースで連続的、選択可能な行動が膨大

・行動選択に用いる「方策関数」自体を学習

ロボット制御などParameterベースでも高い性能を実現

OpenAI Gym



・Brockman et al, 2016

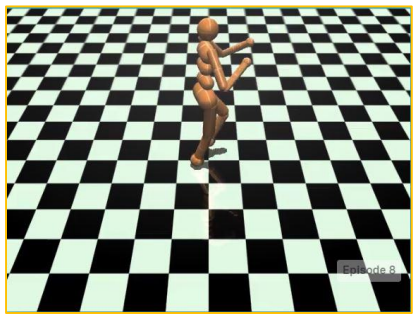
・強化学習のためのシミュレーション用プラットフォーム(Environment)

頻繁に用いられるのがこの3種類、順に難しくなる

・Classic Control (CartPole, MountainCar, etc.)

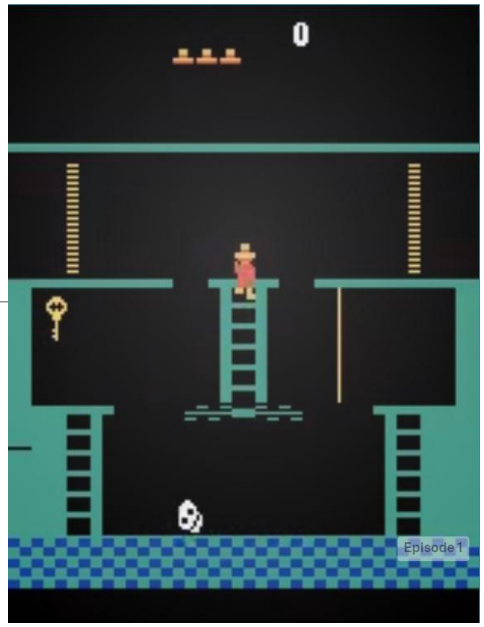
・Atari (The Arcade Learning Environment)

・MuJoCo (Swimmer, Walker2d, etc.)



・上2つはQ学習、MuJoCoは方策勾配法がメイン

Montezuma’s Revenge



・Atariの中で学習が困難とされていた

・報酬が Sparse かつ Delayed

(鍵を取って次のステージへ、障害物に注意)

(報酬は「鍵取得時」と「ステージ移動時」のみ)

・二種類のDQNを用いたh-DQN (Kulkarni et al, NIPS2016)

・状態の新奇性を知るpseudo-count (Bellemare et al, NIPS2016)

・YouTubeのプレイ動画(＋効果音)を活用 (Aytar et al, NIPS2018)

他印象的な研究タスク

①DeepMind Lab (platform, 3D navigation) (Beattie et al, DeepMind 2016)

⑤model に 無相関な external memory を応用 (Oh et al, ICML2016)

②行動探索としてSuprisalを提案 (Achiam and Sastry, 2017)

③DDPG に ME を足し込んだ Soft Actor-Critic (Haarnoja et al, NIPS2017)

④POMDP を階層化で解決した FeUdal Networks (Vezhnevets et al, NIPS2017)

④複数の下位方策から選択する上位方策 (Frans et al, ICLR2018)

③近似誤差の課題推定を解決したTD3 (Fujimoto et al, ICML2018)

①Platform, ②Exploration, ③Algorithm, ④Meta Learning, ⑤Experience Replay

Model, Robot, navigate, NLP, GAN, survey など関連タスクは多岐に渡る

予定とする研究内容

・性能の高い強化学習アルゴリズムへの理解を深める

(TRPO(2015), ACKTR(2017), TD3(2018), IMPALA(2018), 随時収集)

・関連論文における環境構築や再現実験

(一貫したScaleでの比較、Hyper Parameterに因る変動性)

・Parameterベースのタスクでも有効なアルゴリズムの開発

(K-FACを用いて性能を維持しつつ計算時間を減らせないか？)

(行動選択がParameterizeな物理エンジンをタスクとする)

当研究による展望

・環境や条件に適した強化学習アルゴリズムの選定指針を得る

Parameterizeタスクでも適したアルゴリズムは変わってくるはず

・論文同等の性能を得るための指針を得る

比較に用いるScaleやHyper Parameterの選定手順など

・強化学習を用いた技術で短時間化/高性能化が実現する

応用例))ロボット制御、自動運転など