未来画像予測モデルと時間重み付けを導入した価値関数に基づく強化学習

加藤　誉基*，山内　悠嗣（中部大学）

Reinforcement Learning Based on Value Functions Introducing Future Image Prediction Model and Time-weighted

Yoshiki Kato, Yuji Yamauchi (Chubu University)

１．はじめに

強化学習は機械学習の一つであり，自らが行動することで得られる経験から学習する．そのため，学習データを用意することが難しいタスクや未知の環境でもタスクを解くことができる可能性を持つ．強化学習は観測した現在までの状態における価値を最大化するよう学習する．価値とは，将来に亘って獲得できる報酬の期待値であり，西片等は先の状態を予測できれば現在の状態より高い価値を求められるという発想から先の状態を予測するモデルを価値関数に導入した(1)．しかし，現時刻から時間が経過するほど未来の予測は曖昧さを含み不安定となるため，長期の予測を導入した場合，性能が低下する問題を抱えていた．

そこで，本研究ではより高い現在の状態の価値を求めるため，先の状態を予測する際に時間経過に対して重み付けする．これにより，提案手法は直近の未来予測の結果を重視することが可能となり，早期に高い報酬を得ることが期待できる．

２．提案手法

２．１．提案手法の流れ

Fig.1.に提案手法の流れを示す．提案手法は，強化学習パートと未来画像生成パートの 2 つから構成されている．未来画像生成器には学習済みのモデルを使用し，価値を推定する Q ネットワークと行動を決定する方策ネットワークを学習する．

まず，環境から観測した時刻tにおける画像s_tをエンコーダを介して方策ネットワークに入力し，行動a_tを決定する．次に画像s_tと行動a_tを学習済みの未来画像生成器に入力し，1 フレーム先の未来画像$\hat{s}_{t+1}$を生成する．その後，生成した未来画像$\hat{s}_{t+1}$を方策ネットワークに与えたときの時刻$t+1$の行動a_{t+1}を出力し，さらに 1 フレーム先の未来画像$\hat{s}_{t+2}$を生成する．これを繰り返すことでNフレーム先の$\hat{s}_{t+N}$を生成する．最後に生成した時刻$t+N$までの未来画像をエンコーダを介して Q ネットワークに入力し，先の状態の価値$Q(\hat{s}_{t+N}, a_{t+N})$を求め，Q ネットワークと方策ネットワークの重みを更新する．

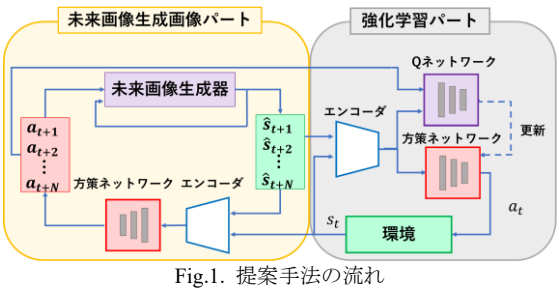


Fig.1. 提案手法の流れ

２．２．未来画像生成

未来画像生成器には Convolutional Dynamic Neural Advection(CDNA)(2)を採用する．CDNA は，連続した数フレームの画像群とそれらの画像群から観測されるオブジェクトの動きや姿勢などを条件として加え，1 フレーム先の未来画像を生成する条件付き未来画像生成器である．

Fig.2.に CDNA の構成を示す．CDNA はエンコーダ・デコーダ型のネットワーク構成であり，エンコーダ，デコーダは畳み込み LSTM をベースとしている．まず，エンコーダに時刻tの画像s_tを入力して得た特徴量と時刻tの条件a_tを結合する．そして，得られた特徴量をデコーダに入力することで，画像s_tに映っているオブジェクトの動きの変化を捉えるフィルタ(移動フィルタ)とオブジェクトの位置を示すマスクフィルタを生成する．その後，画像s_tに移動フィルタを適用し，マスクフィルタとかけ合わせた複合マスクを生成する．最後に画像s_tに複合マスクを適用して 1 フレーム先の未来画像$\hat{s}_{t+1}$を生成する．その後，エンコーダに生成した未来画像$\hat{s}_{t+1}$を与え，時刻$t+1$の行動a_{t+1}を出力し，さらに 1 フレーム先の未来画像$\hat{s}_{t+2}$を生成する．

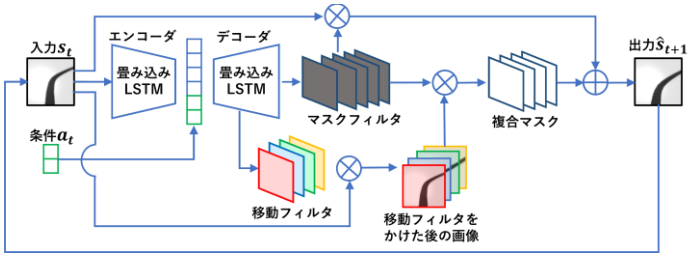


Fig.2. CDNA における未来画像生成の流れ

２．３．強化学習手法

強化学習の手法には，Contrastive Unsupervisd Reinforcement Learning(CURL)(3)を採用する．CURL の観測した画像からランダムに異なる 2 つの領域をトリミングし，anchor と positive の 2 つのデータに拡張する．そして，anchor と positive の潜在変数が類似するようにエンコーダを学習し，Q ネットワークと方策ネットワークで利用する．

従来法(1)における価値関数は式(1)により定義される．

$$L = r_t + \frac{1}{2}\gamma\{Q(s_{t+1}, a_{t+1}) + \frac{1}{N-1}\sum_{n=2}^{N} Q(\hat{s}_{t+n}, a_{t+n})\} - Q(s_t, a_t) \qquad (1)$$

ここで，r_tは時刻tにおいて獲得する報酬，γは報酬の減衰率，$\hat{s}_{t+n}$は未来画像生成器で生成したn時刻先の未来画像を表す．第 1 項は報酬r_t，第 2 項と第 3 項は次時刻からN時刻先の価値$Q(s_{t+1}, a_{t+1})$の和と現在の価値$Q(s_t, a_t)$の差分を損失とし，現在の価値がN時刻先の価値に近づくように学習する．一方，提案手法の損失関数は，時間の経過に対して重みを付けた損失関数として式(2)により定義する．

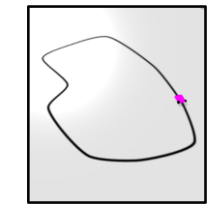
$$L = r_t + \frac{1}{2}\gamma\{Q(s_{t+1}, a_{t+1}) + \sum_{n=2}^{N} \frac{N+1-n}{\sum_{i=1}^{N-1} N-i} Q(\hat{s}_{t+n}, a_{t+n})\} - Q(s_t, a_t) \quad (2)$$

時間が経過するほど未来の予測は曖昧さを含み不安定となるため，提案手法は時間の経過に対して重み付けをする．

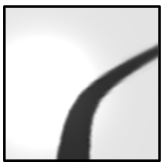
3．評価実験

3．1．実験概要

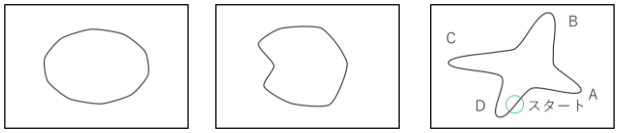
ライントレースタスクを対象に 5 回学習した平均を CURL(3)と従来法(1)，提案手法で比較する．ライントレースタスク環境を Fig.4.(a)に示す．白色の背景に黒いラインのコースをランダムに生成し，エージェントがラインに沿って反時計回りに走行するよう学習させる．エージェントの前方にはカメラが搭載されており，Fig.4.(b)のような画像を撮影し環境の状態とする．エージェントの左右それぞれのタイヤの制御値$[-1.0, 1.0]$をエージェントの行動とし，ラインから逸れないように走行した距離を報酬とする．評価には Fig.4.(c)に示すような難易度の異なる 3 種類のコースを用いて評価する．



(a) ライントレース環境



(b) フロントカメラで観測した画像



(c) 難易度が異なる評価用コースの例

Fig.4.ライントレースタスクの環境と評価用コースの例

3．2．実験評価

Fig.5.(a)に学習中の報酬の遷移，Fig.5.(b)，(c)，(d)に評価コース easy，normal，hard の走行した結果を示す．Fig.5.(a)より 30,000ステップ以降において従来法と提案手法は CURL と比べて高い報酬を獲得できていることを確認することができた．また，Fig.5.(b)，(c)，(d)より全ての評価用コースにおいて提案手法の方が早期に高い報酬を獲得していることが確認できる．

Fig.4.(c)に示す hard コースの A~D までのカーブを走行できた回数を Table1 に示す．なお，各手法において十分な学習ができたと判断できる80,000ステップから100,000ステップまで，5,000ステップ刻みで評価した．Table1 よりカーブを走行できた回数は，多い順に提案手法，従来法，CURL の順になっていることがわかる．CURL に未来画像予測モデルを導入した従来法は，鋭角なカーブの予測が可能となったことからカーブを走行できた回数が増えたと考えられる．また，予測した未来画像の時間経過に対して重み付けした提案手法は，直近の未来予測の結果を重視するため予測の信頼性が向上し，最もカーブを走行できた回数が多かったと考えられる．

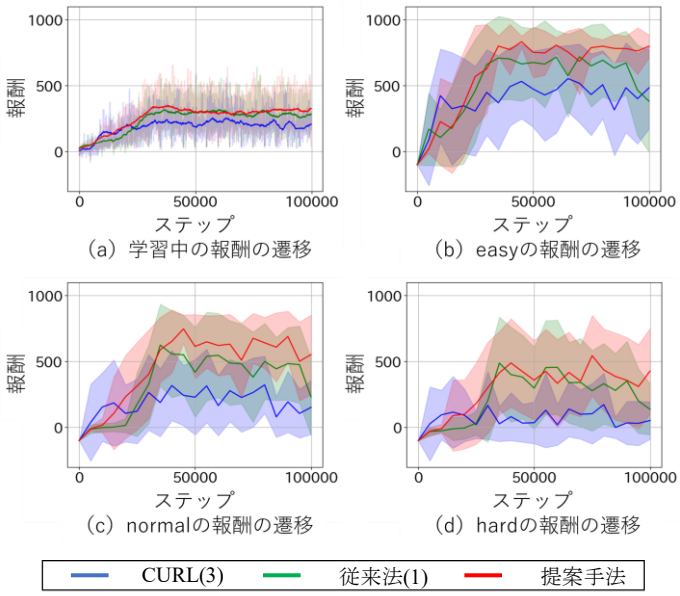


Fig.5. 学習中の報酬と各コースにおける報酬の遷移

Table 1 hard コースにおける A~D のカーブの走行成功回数[回]

手法	A	B	C	D	平均
CURL(3)	0.2	0.2	0.0	0.0	0.10
従来法(1)	1.0	0.8	0.4	0.4	0.65
提案手法	2.6	2.6	2.4	2.0	2.40

4．おわりに

未来画像予測モデルに対して時間重み付けを導入した価値関数に基づく強化学習手法を提案した．提案手法は，予測した未来画像の直近の推定結果を重視するため，ライントレースタスクにおいては安定した走行が可能となった．今後は行動決定時に未来画像予測モデルを導入する手法について検討する予定である．

文　献

(1) 西片 智広，山内 悠嗣，時系列予測モデルを導入した価値関数に基づく強化学習，動的画像処理実利用化ワークショップ，2023.
(2) C. Finn, et al.: Unsupervised learning for physical interaction through video prediction, Advances in neural information processing systems, 2016.
(3) M. Laskin, et al.: Curl: Contrastive unsupervised representations for reinforcement learning, International Conference on Machine Learning, pp.5639-5650, 2020.