Abstract

我々は、強化学習を用いて高次元のセンサー入力から直接制御方針を学習することに成功した初のディープラーニングモデルを提示します。このモデルは、生のピクセルを入力とし、将来の報酬を推定する価値関数を出力とする、Q-ラーニングのバリアントで訓練された畳み込みニューラルネットワークです。我々の方法をアーケード学習環境の7つのアタリ2600ゲームに適用し、アーキテクチャや学習アルゴリズムの調整を行わずに、これまでのアプローチを6つのゲームで上回り、3つのゲームでは人間の専門家を超える性能を達成しました。

Introduction

- 報酬遅延の問題：報酬は行動の直後には得られず、時間が経過してから評価できる。
- シーケンスデータの扱い

Background

- シーケンスデータを扱うことでMDPとして扱える。
- 価値関数を推定したいけどすべての状態と行動の組み合わせに対して価値観数を計算するのは現実的でないのでNNで近似する。NNの訓練はTD誤差（TemporalDifference：予測値と実際の報酬の差）を最小化することで行う。
- 勾配降下法で最適化することで計算コストを削減する。

Related Work

TD-gammonっていうbackgammonのAIが良かったらしい。でも他のゲームにはだめだったらしい。どうやら特化したアプローチみたいだったらしい。具体的にはゲームに使うサイコロが状態空間の探索を助け、価値観数をなめらかにしていたみたい。（なめらかだと収束がしやすい、汎化性能も良い） 他にも問題が見つかって、非線形関数近似とかは収束しないことがあるらしい。ので、線形関数近似を使うことが多かったらしい。 ただ、発散問題は勾配時系列差分法（GTDM）で部分的に解決されたらしい。 NFQっていう先行研究があってバッチ更新のため計算コストにデータサイズが比例する。 本研究は確率的勾配更新を使ってるので大規模データセットに考慮している。

Deep Reinforcement Learning

経験再生の問題について：優先度付き経験再生（Prioritized Experience Reply） そもそも価値関数の更新のために経験再生を行う。

モデル構造の準備

・本研究用の下準備（グレースケールだのなんだの） ・アーキテクチャの提案 従来のNNの使い方は、各アクションのQの値を個別に学習。これはアクション数に比例してしんどい。提案手法では、アクションに対してNNを使うのではなく状態についてNNを使い、その状態がとり得るアクションに対してQの値を出力する。これによりアクション数に依存しない。 フレームスキップ技術：4フレームに1回だけ行動を選択する。これにより計算コストを削減する。要はわざわざ毎フレーム計算し行動選択したところで変わらないことが多いから４フレームに一回でいいってこと。 ・学習と安定性 強化学習は教師あり学習に比べて訓練中のモデルの性能を評価しづらい。理由は方策の重みが変わると移動する方針も大きく変わるので、平均総報酬が安定しないしめっっっちゃノイジー。推定行動価値関数（割引報酬を得ることができるかどうか）は安定して

る。じゃあこれ使えばいいじゃん！？ただ実際に行動をテストしてるわけじゃないから進捗はわかるけど性能を評価できてるわけじゃないのかな？

評価

他の手法との比較 ・SARSA： ・Contingency: 画像の入力のみで強化学習を行っている。これってゲームの中のエージェントの検出ってどうなってるの？←従来は事前知識を与えていたけど、DQNは何もなしに自分で重要な情報を抽出している。すごい。

Conclusion

DQNの適用したらすごかった。