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。