BitTorrent 型並列ダウンロードシステムにおける機械学習を用いた効率的な利得の獲得方法

内藤 世啓† 藤田 聡‡

広島大学 大学院工学研究科 広島大学 大学院工学研究科

1. はじめに

情報通信技術の発展に伴い，P2P 技術に基づく分散システムが広く用いられるようになってきた．その代表例のひとつである BitTorrent[1]では，ダウンロードの対象となるファイルが固定長のピースにあらかじめ分割され，各ピースのアップロード先は，各ピアが局所的に観測した履歴に基づいて分散的に決定される．アップロード先を決定する戦略として「しっぺ返し戦略」が知られているが，しっぺ返し戦略では短期的な履歴しか参照しないため，長期的な観点からは適切に振る舞えなくなる可能性が高い．本稿では，各ピアが局所的に観測した長期的な履歴から，自身の利得を最大化するような戦略を機械学習によって自律的に獲得する手法を提案する．

2. 関連研究

Ratzin ら[2]は，ピースのアップロード先ピアを選択するための適切な「戦略」を強化学習によって獲得する手法を提案した．この手法では隣接ピアからのダウンロード履歴を各ピアの状態として定義し，現在の状態からより多くのダウンロードを受けるために最適と思われる行動(アップロード先となるピアの選択)を強化学習の一種である Q 学習により獲得している．しかしすべての隣接ピアの状態と行動に対する行動価値関数を求めようとすると状態数が爆発し，学習の収束に時間がかかりすぎてしまう．この問題点を解決するため Ratzin らは，学習対象となる隣接ピア数をごく少数に限定するという方法を便宜的にとっている．

また Mnih らは，ビデオゲームにおける最適な行動を学習する方法として，Q 学習に深層学習を適用した DQN(deep Q-network)[3]を提案した．この手法では，Q 学習における行動価値関数をニューラルネットワークを用いて近似することで状態空間をテーブルとして持つことができないほど大規模なゲームに対する Q 学習の適用を可能としている．具体的には DQN では，エージェントがゲームを繰り返しプレイして入力データを生成し，それらのデータから Q Network と呼ばれる畳み込みニューラルネットワーク(CNN)を最適化していく．Q Network は状態$s \in S$が入力データとして与えられたとき，状態sでエージェントが行動$a \in A$を行った場合の行動価値関数$Q(s,a;\theta)$を出力する．ここでθは CNN の重みである．

DQN では，学習を安定して収束させるための工夫がいくつかなされている．一つ目は過去の遷移$e_t = (s_t, a_t, r_t, s_{t+1})$を Replay Memory と呼ばれるバッファに格納しておき，学習の際には Replay Memory からランダムサンプリングして θ を更新する方法である．もしランダムサンプリングを行わず，エージェントが観測した順に学習をさせてしまうと，データ間の相関が高くなり過ぎ，学習が不安定になりやすい．二つ目は，Q Network の教師信号の Q 値を出力するための Target Network を作り，Q Network のパラメータを定期的に Target Network にコピーして次の更新まで固定する方法である．これにより，教師信号と予測するQ関数の相関を小さくし，学習を安定させることができる．

3. システムモデル

BitTorrent を単純化した次のようなシステムモデルを用いる．完全結合された均一なピアの集合$\boldsymbol{P} = \{p_0, p_1, \ldots, p_{N-1}\}$を考える．以下ではダウンロードを要するピアをリーチャ，すべてのピースのダウンロードを完了したピアをシーダとそれぞれ呼ぶ．各ピアの基本動作は次の通り：

- 時刻tで各ピアは，隣接リーチャの中からアップロード先ピアを高々1 台選択する．
- リーチャは，他ピアからアップロード先として合計 x回以上選択されると十分なアップロードを得られたものとしてシーダとなる．
- シーダはシステム内に残り続け，アップロ

Learning of efficient strategy for peers participating in parallel download systems of BitTorrent type

† Tokihiro NAITO, Hiroshima University

‡ Satoshi FUJITA, Hiroshima University

ードのみを行う．シーダによるアップロード先の選択はランダムに行われる．

以下では，時刻tにおける隣接ピアからのダウンロードの様子と隣接ピアへのアップロードの様子をそれぞれ長さ$\mathrm{N}-1$のビットベクトルで表現する(ピースの送信があった場合は1，なかった場合は0とする)．

4. 提案手法

DQN によってアップロード戦略を獲得させるピアをp_0に固定する．p_0がリーチャとしてシステムに参加してからシーダになるまでを 1 回のエピソードとし，そのようなエピソードを繰り返し実行する．p_0以外のピアは，エピソード毎にアップロード戦略を変化させても構わない(変化方法は後述)．このようなエピソードを何度も経ることによって，(環境が動的に変化していく中でも)より多くのダウンロードを獲得することのできる戦略をp_0に学習させることを目指す．

提案手法で用いる状態，行動，報酬の定義は以下の通りである：時刻tにおける状態s_tは，ピアp_0が観測した時刻$\mathrm{t}-h$から$\mathrm{t}-1$までのアップロード・ダウンロードの履歴であり，行動a_tは，p_0が時刻tにとった行動である．ここで行動は，1 台の隣接ピアを選択するかどれも選択しないかの計N通りの中から選択される．また報酬r_tは，行動a_tを行った後に対象となる隣接ピアから受けた被選択回数である．状態sは Q Network の入力データでもあるため，学習を容易にするために，各列が各時刻におけるアップロード・ダウンロード状況であるような二次元ビット配列として整形する．このように関連性が高い数値を近い位置に配置しておくことによって CNN による特徴抽出の方法を適用することができる．1 回のエピソードの中でp_0による行動の選択は毎時刻行われる．具体的には，1) 確率ϵで N 通りの選択肢の中からの選択をランダムに行い，2) 確率$1-\epsilon$で Q Network への問い合わせを行い，入力データs_tを Q Network に与えたときの出力$Q(s_t,a;\theta)$において最も Q 値の高い行動を決定的に選択する．また行動選択後，行動a_tによって得られた報酬r_tと次の状態$\mathrm{s}_{\mathrm{t}+1}$を観測し，それらの結果を Replay Memory に組(s_t,a_t,r_t,s_{t+1})として保存する．

学習中の Q Network の更新方法は以下の通りである：あらかじめ決められたミニバッチサイズ分だけ Replay Memory からランダムサンプリングし，サンプリングデータ(s_j,a_j,r_j,s_{j+1})と Target Network を用いて Q Network の教師信号y_jを作成する．具体的には，もし時刻$\mathrm{j}+1$でそのエピソードが終了する場合は$\mathrm{y}_\mathrm{j}:=r_j$とし，そうでない場合は $\mathrm{y}_\mathrm{j}:=r_j+\gamma\max_{a'}\hat{Q}(s_{j+1},a';\theta^-)$ とする．ここで$\hat{Q}(s_{j+1},a';\theta^-)$は状態$\mathrm{s}_{\mathrm{j}+1}$を入力データとして与えた時の Target Network の出力であり，その最大値をとることで，最適な行動を将来的にとり続けた場合の累積報酬の期待値を予測している．ここで$\gamma(0<\gamma<1)$は将来予測における割引率である．

5. 評価

今回の実験では，提案手法で得られたアップロード戦略と(a) ランダム戦略，(b)しっぺ返し戦略，(c)しっぺ返し＋楽観的アンチョーク戦略で行動を決定した場合のシーダになるまでにかかる時間の比較を行った．(a)はランダムに行動を選択し，(b)は直近にアップロードを受けた隣接ピアを選択するが，誰からもアップロード受けなければアップロードしない選択をする．(c)は直近にアップロードを受けた隣接ピアを選択し，誰からもアップロード受けなかった場合は，ランダムにアップロード先を選択する．

エージェントの学習環境は$\mathrm{N}=10$，$\mathrm{x}=100$，$\mathrm{h}=20$，$\gamma=0.9$とし，ϵは最初の100万回の行動選択の間に，1.0から0.1まで線形に減少させ，その後0.1に固定する．エージェント以外の各ピアは，(a)，(b)，(c)の中からエピソード毎にランダムに戦略を変化させ，計2万回のエピソードを学習させ，提案手法の戦略を獲得する．

同様の環境でピアp_0が各戦略をとり，それぞれ3万回の試行を行い，シーダになるまでにかかる時間を比較した．その結果，提案手法では90.6%の割合で最も早くシーダになることが分かった．その他の戦略では(a)0.2%，(b)0.007%，(c)9.2%となっている．提案手法で得られた戦略は，様々な環境において他の戦略よりダウンロードを獲得できている．

6. おわりに

今後の課題として，より汎用性の高い戦略の獲得方法を確立することなどがあげられる．

参考文献

[1] "BitTorrent," http://www.bittorrent.com.

[2] R. Ratzin, et al., "Online Learning in BitTorrent Systems, " IEEE Trans. Parallel Ditrib. Syst., 23(12): 2280-2288, Mar. 2012.

[3] V. Mnih, et al., “Human-level control through deep reinforcement learning, ” Nature, 518: 529-533, 2015.