GPGPU を用いたオンライン機械学習の高速化

2012SE087 加藤大貴
指導教員：宮澤元

1 はじめに

機械学習は，商品レコメンデーションやパターン認識など，大量のデータの分析に利用されている．これらに用いられている機械学習はバッチ学習と呼ばれる，溜め込んだデータに対して一度だけ学習を行うものである．大量のデータをバッチ学習によって学習する場合，計算量が大きいため，General-purpose computing on GPU(GPGPU) [1] による並列化が行われている．

一方，ソーシャルネットワーキングサービスやスマート家電などの登場によって大量のデータがリアルタイムに発生している．これらのデータを分析するために，オンライン機械学習と呼ばれる機械学習のアルゴリズムが提案されている．オンライン機械学習とは，連続して発生するデータに対して学習を繰り返し，学習モデルを常に更新し続ける機械学習である．オンライン機械学習を行うために，Jubatus [2] などのフレームワークが開発されている．GPGPU によってオンライン機械学習を並列化する場合，連続して発生するデータに対して学習を繰り返すという特徴から，ホストメモリとデバイスメモリ間でのデータコピーの回数が多くなり，高速で処理を行うことは困難である．

本研究の目的は，GPGPU を用いてオンライン機械学習を高速化することである．オンライン機械学習を GPGPU によって高速化することは，データコピーによるオーバヘッドが大きいため困難であるが，データの発生パターン等の条件によっては高速化が可能と考える．

そこで，本稿では，Jubatus のソースコードを基にクラスタリングアルゴリズムの一つである k 平均法に GPGPU を適用し，様々な条件の下で実際にオンライン機械学習を行い，GPGPU を適用することによって処理の速度が向上するオンライン機械学習の条件について考察を行う．また，実際に考えうるシチュエーションとして，カメラから流れ込むデータの k 平均法による圧縮を想定した実験を行い，実験で得られた GPGPU の適用が有効な条件が正しいことを確認する．

2 研究背景

現在，ビッグデータを処理するためにオンライン機械学習や GPGPU といった技術が利用されている．本章ではそれらの技術について説明を行う．

2.1 オンライン機械学習

オンライン機械学習とは，次々と到着するデータに対してその場で学習を行い，学習結果を随時更新することのできる機械学習のことである．オンライン機械学習に対して，溜め込んだデータに対して学習を行い，その後は学習結果を更新しない一般的な機械学習をバッチ学習と呼ぶ．オンライン機械学習の利点は，新たなデータが発生したとき，直ちにそのデータを学習結果に反映できる点である．また，バッチ学習では新たなデータが追加された際，追加されたデータと追加される前のデータを合わせて再び学習し直す必要があり，多くの計算時間が必要である．オンライン機械学習向け分散処理フレームワークの 1 つとして，Preferred Networks と NTT ソフトウェアが開発した Jubatus がある．

2.2 GPGPU

ビッグデータを高速に処理する手段の 1 つとして，GPGPU が挙げられる．GPGPU は，Graphics Processing Unit(GPU) を汎用計算に用いるための技術である．画像処理を行うためのもである GPU は，マルチスレッドによる並列計算を行う能力が CPU より優れている．

GPGPU で処理を行うためには，ホストメモリからデバイスメモリにデータをコピーする必要があるが，この処理には時間がかかり，オーバヘッドとなるので，GPU による処理を効率よく行うためにはデータコピーの回数を極力少なくする必要がある．

2.3 GPGPU のオンライン機械学習への適用

2.2 節で述べたように，GPGPU で効率よく処理を行うためにはホストメモリとデバイスメモリ間でのデータコピーの回数を極力少なくする必要がある．しかし，オンライン機械学習は常に発生しつづけるデータに対して学習を行う機械学習であるので，データコピーの回数は必然的に多くなる．従って，オンライン機械学習は GPGPU による並列化に向かない処理であるといえる．しかし並列処理を行うことによって短縮される処理時間がデータコピーに必要な処理時間を上回る場合など，条件によって GPGPU を適用することによって処理時間を短縮することができる場合もあると考えられる．

2.4 関連研究

オンライン機械学習に GPGPU を適用し，高速化を行った研究に “Fast On-line Statistical Learning on a GPGPU” [3] がある．この研究は，オンライン機械学習の手法，確率的勾配降下法を GPGPU を用いて高速化した “Delayed Stochastic Gradient Descent” [3] を提案している．

通常の確率的勾配法では，学習モデルを更新する時に直前のデータを反映させたモデルが必要であり，並列化を行うことができないが，“Delayed Stochastic Gradient

Descent” では，直前のデータを反映したモデルを用いず，いくつか前のデータを反映したモデルを用いて学習モデルを更新することによって並列化を行っている．この手法では，より前のモデルを用いることによって並列度を上昇させることができるが，学習の精度が下がってしまう．

3 実験

オンライン機械学習に GPGPU を用いることによる処理時間及び処理結果の変化を計測するための実験を行った．実験では，機械学習の手法のうち，k 平均法と呼ばれる手法を用いて GPGPU 適用前後の処理時間を計測した．k 平均法のうち，繰り返し処理によってデータとクラスタの距離の計算を行っていた部分を GPGPU によって並列化した．本章では，行った実験の内容と結果について述べる．

3.1 実験の概要

k 平均法の以下のパラメータを変化させ，それぞれ処理時間及び処理結果を計測した．

- compressed_bucket_size
- k
- bucket_size
- bucket_length

上の 4 つのパラメータに加え，データ数とデータの送信回数を変化させ，実験を行った．

3.2 結果

データ数のみを変化させた場合の実験結果について述べる．データ数 250 の時は GPGPU 適用前が 1.134 秒，GPGPU 適用後が 3.110 秒と GPGPU 適用後の方が処理時間が長かったが，データ数が増えるとともに GPGPU 適用後の処理時間が GPGPU 適用前に対して短くなり，データ数 2250 の時に GPGPU 適用前が 29.894 秒，GPGPU 適用後が 30.668 秒とほぼ同じになった．データ数 6250 の時，GPGPU 適用前が 229.316 秒，GPGPU 適用後が 134.766 秒と，GPGPU 適用後の処理時間の方が大幅に短くなった．

3.3 実験から分かったこと

実験の結果，特徴ベクトルの次元，データ数は多いほど，圧縮率は低いほど GPGPU 適用後の処理時間が適用前より短くなること，クラスタリング回数，データの送信回数，クラスタ数が処理時間に及ぼす影響は GPGPU 適用前後でほぼ変わらないことが分かった．

3.4 考察

実験の結果から考えられる，GPGPU の適用が有効な機械学習の他の例として，近傍探索の一つである MinHash が挙げられる．MinHash には，与えられたデータのハッシュ値を計算する処理があり，この処理はデータの数だけ繰り返される．そのため，実験と同様に GPGPU による並列化を行うことで，高速化が可能であると考えられる．

4 GPGPU の適用が有効な機械学習の具体例

本章では，第 3 章で得られたオンライン機械学習の各パラメータと処理時間の関係から，GPGPU の適用が特に有効であるオンライン機械学習の具体例を挙げ，それについての実験と考察を行う．GPGPU の適用が特に有効であるオンライン機械学習の例として，カメラから流れ込む画像データの k 平均法による圧縮が考えられる．

4.1 画像圧縮の実験の概要

5 秒毎にカメラから流れ込むデータに対してクラスタリングを行う処理を，GPGPU 適用前後のそれぞれのプログラムで 10 回ずつ実行し，処理結果と処理時間を比較した．

4.2 画像圧縮の実験の結果

計測の結果，クラスタリングの処理結果は GPGPU の適用の有無に関わらず同じであった．また，10 回目のクラスタリングを終えるために要した時間は，GPGPU 適用前のプログラムが 137 秒，GPGPU 適用後のプログラムが 53 秒であった．これらのことから，カメラ画像の k 平均法による圧縮という処理に対して GPGPU を適用することで学習結果の質を保ちつつ学習速度を向上させることができるといえる．

5 おわりに

Jubatus の k 平均法プログラムを用いて，オンライン機械学習に GPGPU を適用した場合の処理時間の変化についての実験を行った．その結果，データ数が多い，処理回数が多いといった幾つかの条件下において GPGPU の適用が有効であることがわかった．また，画像圧縮をシミュレートした実験によって，実際の問題に対しても GPGPU の適用が有効であることが確かめられた．

本稿では，オンライン機械学習に GPGPU を適用した場合の処理時間の変化を調べるために k 平均法を使用して実験を行ったが，他のオンライン機械学習の手法に対して GPGPU を適用した場合の効果が同じであるとは限らず，今後，k 平均法以外の手法についても実験を行う必要がある．

参考文献

[1] GDEP `http://www.gdep.jp/page/view/248`(2016 年 1 月 7 日閲覧)

[2] Jubatus `http://jubat.us/ja/`(2016 年 1 月 7 日閲覧)

[3] Xiao, F., McCreath, E. and Webers, C. (2011). Fast On-line Statistical Learning on a GPGPU. In Proc. Australasian Symposium on Parallel and Distributed Computing (AusPDC 2011), Perth, Australia. CRPIT, 118. Jinjun Chen and Rajiv Ranjan Eds., ACS. 35-44.