

GPU によるニューラルネットワークの高速化

D-2

Acceleration of Neural Network by GPU

橋本 裕介[†]Yusuke HASHIMOTO[†][†] 東京都立産業技術高等専門学校[†] Tokyo Metropolitan College of Industrial Technology黒木 啓之[†]Takashi KUROKI[†]

1. はじめに

近年 GPGPU を用いたニューラルネットワークに関し、非常に多くの研究が行われている。高田ら[1]のように1ブロックあたりのスレッド数による計算比較を行う研究は多いが、ブロック数による計算比較を行なっている研究は少ない。そこで本研究ではニューラルネットワークの学習計算を対象とし、スレッド数とブロック数の増減によって検証を行い、並列計算の高速化を目的とする。

2. GPGPU

GPU は、一般的に画像処理を専門とする演算装置である。画像処理が高負荷ながらも簡単な演算の繰り返しであることから、多数の演算コアを利用しこれらの演算を並列処理することに長けている。GPGPU は、この GPU の演算能力を画像処理以外の目的に応用する技術のことである。

3. CUDA

CUDA は、NVIDIA が開発及び提供を行っている GPU 向けの総合開発環境であり、NVIDIA 製のハードウェア性能を最大限引き出すことができるように設計されている。ライブラリ言語である他の GPGPU 向け開発環境に比べ、NVIDIA 製という制約があるものの並列実行の最小単位のカーネルを呼び出す手段が通常のプログラムに似た記述で行うことができ、本来の実装に近い形でプログラムを生成できる。

4. 実験

本稿では、CUDA を利用する GPU に適応した多数決問題を行うプログラムを作成し、スレッド数、ブロック数による並列化を行い、その計測結果からそれぞれの並列化による計測時間の変化について考察を行う。

4.1 実験環境

表 1 に本研究で使用した環境を示す。

表 1 実験環境

GPU	Geforce GTX 680 (コア数:1536)
OS	CentOS 6.7
CUDA	V7.0.27

4.2 実験内容

3 層のニューラルネットワークにおける 3 入力多数決問題に対して計算の並列化を行い、並列数を 1 から 5 まで 1 ずつ増加させ、並列数によって実行時間にどのような変化が現れるのかについて計測を行った。実験の結果を図 1 に示す。横軸にスレッド、ブロックでの並列数、縦軸に GPU で計

算された時間を示す図より、スレッドによる並列化とブロックによる並列化による変化はこの段階では存在しないことがわかった。また、並列数を増やすと時間は減少傾向にあるが、ただ増やせば減少するというわけではないこともわかった。

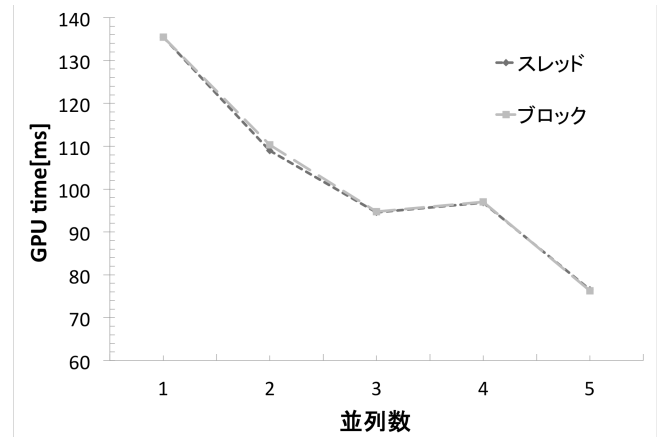


図 1 多数決問題に対する並列数と計測時間の関係

5. 考察

図 1 より、並列数を増やしても時間に変化が起こらない箇所があった。並列数を増やせばスレッド毎での計算時間は減少するにも関わらず、スレッド数を 3 から 4 に変化した時に時間変化が見られなかった。この先は解明する必要がある。

6. 結論

今回の計測結果では、並列数を増やすことによって時間変化は現れるがむやみに並列箇所を増やしても時間が短縮されるわけではないこと、スレッドによる並列化とブロックによる並列化には変化が現れないことがわかった。

今後は実験から得られた疑問点に対し理由を解明し、より並列数の多い実用を想定したプログラムの作成を行い、速度を計測することによって高速化に繋がる法則性を見つける。

参考文献

[1]高田 直樹, 下馬場 朋禄, 増田 信之, 伊藤 智義
シングル CUDA 互換 GPU を用いた FDTD 法の計算高速化 FIT2010(第 9 回情報科学技術フォーラム) C-009, 2010. 9.