平行程式設計實務期末專題報告

陳風平

國立台北科技大學　資訊工程系

110 年 06 月 18 日

目錄

研究目的

- 加速本人類神經網路 MATLAB 專案[1]（以下簡稱本專案）。
- 學習如何加速類神經網路訓練。
- 學習 MATLAB 平行與 GPU 函式庫。

[1] https://github.com/phogbinh/handwritten-digit-recognition

第一階段：探索

- 參考 Block Multiplication 加速方法 [3]。
- 參考 [7] – 用 C++/CUDA 在 GPU 做 convolution 的計算[2]。
- 學習 MATLAB 平行與 GPU 計算教學系列 [1]。

[2] https://code.google.com/archive/p/cuda-convnet

第二階段：初步啓發

- 整理原本程式碼 – 類別、函式、命名等。
- Brainstorm 加速方法與問題（第 6 頁）。
- 決定專題策略：從實驗啓發。

第二階段：加速方法與問題

方法	問題
把全部資料搬到 GPU 做計算	本專案最大矩陣爲 $w^2_{47\times784}$，無法利用 GPU 矩陣相乘加速 [5]
寫 C++ single precision 矩陣相乘 link 到 MATLAB 加速	MATLAB 本身 BLAS 矩陣相乘已 highly-optimized[10, 9]，要花出很多功夫才能跟它速度相比
用 C++ 重寫本專案	要處理龐大資料儲存在記憶體裡面，要研究 C++ 線性代數圖書庫（如 Eigen3、GMTL 等 [4]）

第三階段：實驗

- 使用 MATLAB Profiler 查看程式瓶頸 [6]。
- 參考 MATLAB 提升效率建議 [8]。
- 標註本專案可改做平行的部分[3]。

[3] https://phogbinh.github.io/handwritten-digit-recognition/train_fast.m

第三階段：實驗結果

方法	加速時間（妙）
把 layer 屬性改變數，解開迴圈	250
把 layer_associates 屬性改變數	150
在每個 mini-batch 用 parfor 平行	失敗[4]
重用已配置記憶體的變數 [2]	1
取代全部全域變數	50
把全部資料搬到 GPU 做計算	失敗[5]

[4]跑了 25 分鐘還沒訓練完成。
[5]跑了 40 分鐘還沒訓練完成。

研究成果

- Demo 完整版[6]。
- 成功把本專案加速了兩倍（原 911.8774 秒變 457.9892 秒）。
- 觀察到 MATLAB 本身有 multi-threading（原 CPU 使用率 39% 變 60%）。

[6] https://youtu.be/a7IcN0bq5Z8

參考文獻

[1] Harald Brunnhofer. *Parallel and GPU Computing Tutorials*. 2015. URL: https://www.mathworks.com/videos/series/parallel-and-gpu-computing-tutorials-97719.html.

[2] Stephen Cobeldick. *Setting array elements to zero, the best way*. 2019. URL: https://www.mathworks.com/matlabcentral/answers/443220-setting-array-elements-to-zero-the-best-way.

[3] CodeEmporium. *How do GPUs speed up Neural Network training?* 2020. URL: https://youtu.be/EKD1kEMNeeU.

[4] Reed Copsey and Catskul. *What are the most widely used C++ vector/matrix math/linear algebra libraries, and their cost and benefit tradeoffs?* 2009. URL: https://stackoverflow.com/questions/1380371/what-are-the-most-widely-used-c-vector-matrix-math-linear-algebra-libraries-a.

[5] Jason Dsouza. *What is a GPU and do you need one in Deep Learning?* 2020. URL: https://towardsdatascience.com/what-is-a-gpu-and-do-you-need-one-in-deep-learning-718b9597aa0d.

[6] Doug Hull. *Profiler to Find Code Bottlenecks*. 2006. URL: https://www.mathworks.com/videos/profiler-to-find-code-bottlenecks-97502.html.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90. DOI: https://doi.org/10.1145/3065386.

[8] MathWorks. *Techniques to Improve Performance*. 2021. URL: https://www.mathworks.com/help/matlab/matlab_prog/techniques-for-improving-performance.html.

[9] Cleve Moler. *MATLAB Incorporates LAPACK*. 2000. URL: https://www.mathworks.com/company/newsletters/articles/matlab-incorporates-lapack.html.

[10] James Tursa. *Speed of Matrix-Multiplication (in Matlab, C, other PCs)*. 2015. URL: https://www.mathworks.com/matlabcentral/answers/235094-speed-of-matrix-multiplication-in-matlab-c-other-pcs.