平行程式設計實務期末專題報告

陳風平[1]

資訊工程系、國立台北科技大學、台北

110 年 06 月 18 日

[1]學號：106590048

概要

我們透過實驗加速兩倍一份 MNIST 前饋神經網路的 MATLAB 程式。

目錄

Chapter 1

研究動機

近期的深度學習，不管是在學術界還是工業界，都是熱度頗高的技術之一。而大多數的深度學習模型都使用 NVidia 的 GPU，早就已經不拿 CPU 來做訓練了。

剛好我去年有自己寫出一份辨識手寫數字達 97% 精準度的類神經網路模型[1]（一下簡稱本專案[2]）。但那時候每次在 CPU 執行都要花接近半個小時（訓練與測試），造成我研究上的困擾。因此，我想針對這次平行程式設計實務期末專題（一下簡稱本專題）應用 CUDA 來加速此模型。

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

[2] 本專案爲單純使用矩陣乘法、無用任何其他外掛的 MATLAB 專案。

Chapter 2

研究目的

我眞正的目的爲研究電腦科學家如何加速類神經網路。

因此，我打算先自己想如何優化本專案，再去參考他人的方法。最後，我會選出一個實作又簡單、效果又不錯的方法來改進我的機器學習模型。這樣我不只學會他人的方法，也能接觸 MATLAB 平行與 GPU 函式庫，對我未來上碩士做學術研究有非常大的幫助。

Chapter 3

平行方法

如本專案簡報的研究過程[1]指出，此類神經網路無法做平行計算。我會在此章節帶過該簡報的重點。

3.1 問題

下方表格爲本專案加速方法與問題：

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 [3]
寫 C++ single precision 矩陣相乘 link 到 MATLAB 加速	MATLAB 本身 BLAS 矩陣相乘已 highly-optimized[5, 4]，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 [2]）

Table 3.1: 加速方法與問題

[1] https://phogbinh.github.io/NTUT2021SpringCUDA/final_project/presentation/presentation.pdf

3.2 方法

雖然本專案無法做平行計算，但透過實驗，我成功地把模型訓練的部分加速了兩倍：

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[2]
重用已配置記憶體的變數 [1]	1
取代全部全域變數	50
把全部資料搬到 GPU 做計算	失敗[3]

Table 3.2: 實驗結果

Chapter 4

研究成果

我成功地把本專案模型訓練的部分加速了兩倍（原 911.8774 秒變 457.9892 秒[1]）– 達標 proposal 預期結果。

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

Chapter 5

結論心得

透過本專題，我不只學會了 GPU 在類神經網路加速龐大矩陣乘法的用途，也上手了 MATLAB 平行與 GPU 函式庫。另外，我也順便學習到如何使用 LaTeX 的 beamer 套件做簡報，非常有成就感。

Bibliography

[1] 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.

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

[3] 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.

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

[5] 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.