2020「中技社科技獎學金」

2020 CTCI Foundation Science and Technology Scholarship

研究獎學金

Research Scholarship

基於非揮發性記憶體系統之實現類神經網路的高效訓練與無損精準度協同損耗式寫入

Achieving Lossless Accuracy with Lossy Programming for Efficient Neural-Network Training on NVM-Based Systems

國立臺灣大學資訊工程學研究所 博士班四年級 王韋程

指導教授：郭大維教授、張原豪研究員

國立臺灣大學 National Taiwan University

研究重點

Neural networks over conventional computing platforms are heavily restricted by the data volume and performance concerns. While non-volatile memory offers potential solutions to data volume issues, challenges must be faced over performance issues, especially with asymmetric read and write performance. Beside that, critical concerns over endurance must also be resolved before non-volatile memory could be used in reality for neural networks. This work addresses the performance and endurance concerns altogether by proposing a data-aware programming scheme. We propose to consider neural network training jointly with respect to the data-flow and data-content points of view. In particular, methodologies with approximate results over Dual-SET operations were presented. Encouraging results were observed through a series of experiments, where great efficiency and lifetime enhancement is seen without sacrificing the result accuracy.

研究成果

Why NVM-Based Systems for NN Training?

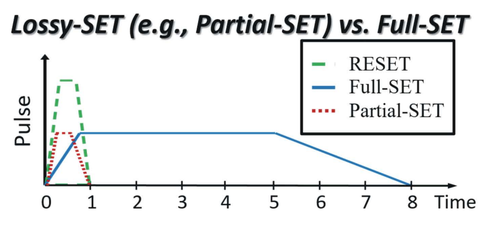
❖ **Dilemma of Big Data and Neural-Network Training on DRAM:**

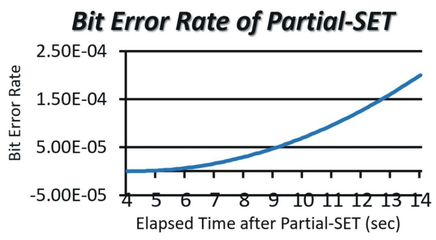
	VGGNet-D	ResNet	GoogLeNet
Input Data	ImageNet (224x224x3)		
Mini-batch Size	256	256	1024
Size of Intermediate Data	**15 GB**	**54 GB**	**24 GB**
Size of Weights	**528 MB**	**230 MB**	**51 MB**
Size of Biases	**52 KB**	**71 KB**	**8 KB**

❖ **NVM-Based (PCM-Based) Training Solutions:**

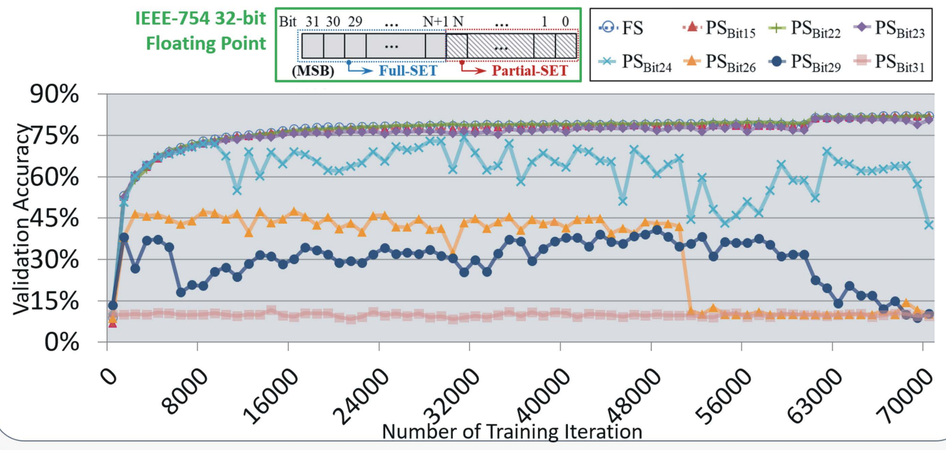
+ Density, Unit Cost, Endurance, Read Latency, Nearly-Zero Leakage Power
- Write Latency (Asymmetric Write), Wear Out Problem

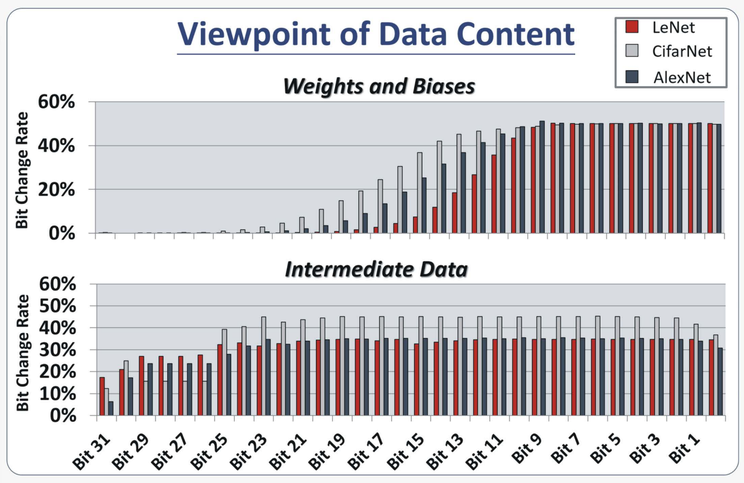
✓ **Lossy Programming (Lossy-SET)**

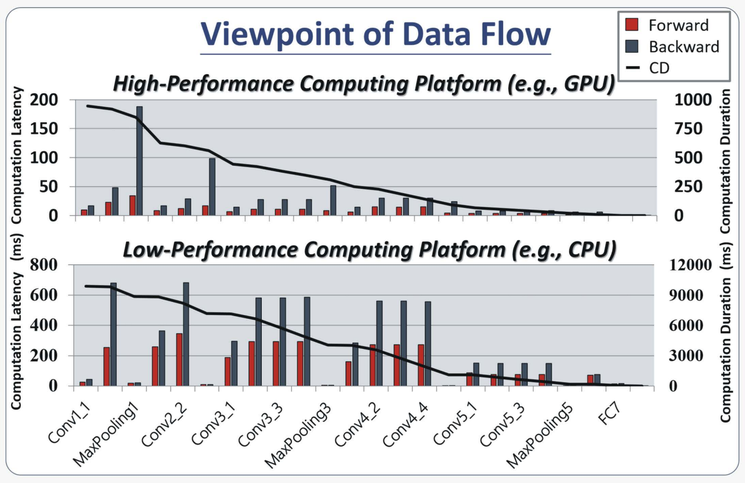




❖ **How to Explore a Great Programming Scheme with the Considerations of Performance, Retention Time, Endurance of NVM and Accuracy of NN**



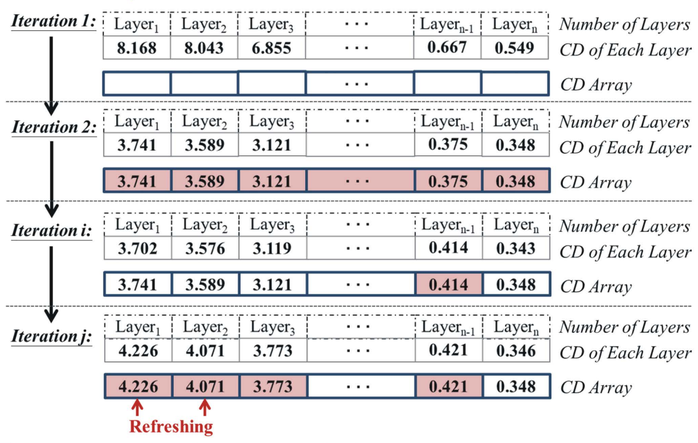




Data-Aware Programming Design (DAP)

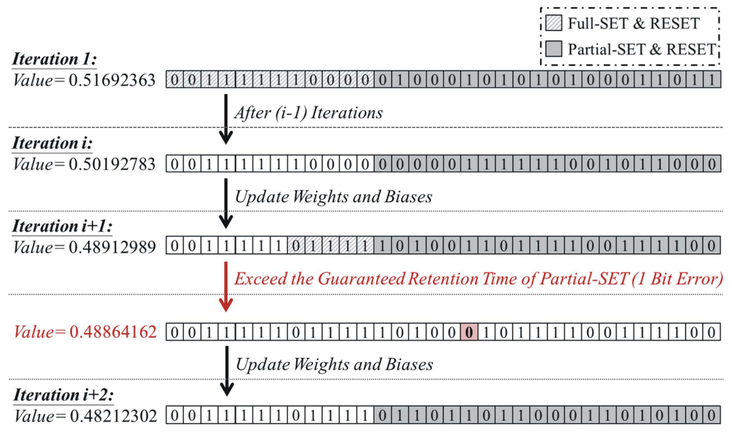
❖ **Layer-Aware SET Policy:**

- Viewpoint of Data Flow - Intermediate Data



❖ **Bit-Aware Dual-SET Policy:**

- Viewpoint of Data Content - Weights and Biases



❑ **Experiment Results:** Remarkably improve the average memory access latency up to 4.3x and enhance the lifetime of PCM to 3.4x

	LeNet			CifarNet			AlexNet			GoogLeNet			VGGNet-D		
	Baseline	PS	DAP	Baseline	PS	DAP	Baseline	PS	DAP	Baseline	PS	DAP	Baseline	PS	DAP
Performance	1x		4.34x	1x		4.24x	1x		4.26x	1x		4.33x	1x		4.26x
Top-1 Accuracy	99.11%	10.18%	99.13%	82.02%	10.52%	81.61%	56.52%	0.07%	56.42%	62.44%	0.09%	62.56%	66.92%	0.1%	66.76%

研究生活及心得

本人目前為就讀於國立台灣大學資訊工程學研究所博士班四年級的博士候選人，致力於精進電腦資訊領域的研究能力；本人自碩士班起，便加入郭大維教授的嵌入式系統暨無線網路實驗室，鑽研於記憶體/儲存系統領域，至今累積了不錯的研究成果，多篇論文皆發表於國際頂尖學術期刊及研討會，而本次所獲獎的研究成果更是榮獲ACM/IEEE CODES+ISSS 2019最佳論文獎，為該會議28年來首度由台灣研究團隊獲得此獎。

在就讀碩、博士班這段期間，本人必須特別致謝郭大維教授及張原豪教授孜孜不倦的教導，使本人除了在學術專業技能及外語能力上有所進步之外，更強烈地感受到自己心境上的轉變，例如正向思考與謙卑心態皆是我從中深刻學習到的道理。因此，無論我將來畢業後是進入業界或學界、在國內抑或是國外，我必定都會抱持著在研究所期間所訓練出的能力與心態，勇敢地跨出未來每一步。