합성곱 신경망 성능 향상을 위한 메모리 내 연산 구조

정건모 [1], 염호윤 [1], 김한준 [2]
[1] 연세대학교 전기전자공학과 통합과정
[2] 연세대학교 전기전자공학과 교수

kunmo@yonsei.ac.kr, hoyun@yonsei.ac.kr, hanjun@yonsei.ac.kr

Processing-in-Memory Architecture for Enhanced Convolutional Neural Network Performance

Kun-Mo Jeong[1], Ho-Yun Youm[1], Han-Jun Kim[1]
[1]Dept. of Electrical and Electronic Engineering, Yonsei University

요 약

최근 고성능 컴퓨팅 장치의 수요 증가와 함께, 메모리 내에 연산을 가능하게 하는 하드웨어 구조가 새로이 발표되고 있다. 본 논문은 기존 DRAM 에 계산 유닛을 통합하는 새로운 메모리 내 연산 구조를 제안한다. 특히, 데이터 집약적인 합성곱 신경망 작업을 위해 최적화된 이 구조는 기존 메모리 구조를 사용하면서도 기존 구조에 분기를 추가함으로서 CNN 연산의 속도와 에너지 효율을 향상시킨다. VGG19, AlexNet, ResNet-50 과 같은 다양한 CNN 모델을 활용한 실험 결과, PINN 아키텍처는 기존 연구에 비해 최대 2.95 배까지의 성능 향상을 달성할 수 있음을 확인하였다. 이러한 결과는 PINN 기술이 저장 및 연산 성능의 한계를 극복하고, 머신 러닝과 같은 고급 어플리케이션의 요구를 충족시킬 수 있는 방안임을 시사한다.

1. 서론

합성곱 신경망(Convolutional Neural Network, CNN)의 활용도가 급증함에 따라, 이를 지원하기 위한 고성능 컴퓨팅 시스템의 필요성이 커지고 있다. CNN 은 자동차 자율 주행, 의료 이미지 분석, 이미지 및 비디오 분류 등 다양한 분야에서 중요한 역할을 수행하고 있으며, 이러한 기술들은 대량의 데이터를 신속하게 처리할 수 있는 능력을 요구한다. 그러나 기존의 컴퓨팅 시스템은 메모리 대역폭의 한계와 높은 전력 소비 문제로 인해 이러한 요구를 충분히 만족시키지 못하고 있다.

이 문제를 해결하기 위해, 본 논문에서는 새로운 메모리 내 연산(Processing In Memory, PIM) 구조인 PINN 을 제안한다. PINN 은 기존 DRAM 을 활용하면서도 CNN 연산의 속도와 효율을 혁신적으로 향상시킬 수 있는 구조를 제공한다. SK 하이닉스의 'Newton' [1] 아키텍처에서 영감을 받아 개발된 이 아키텍처는 각 메모리 뱅크에 독립적으로 배치된 계산 유닛을 통해 데이터 처리를 병렬로 수행함으로써, 복잡한 데이터 연산을 더욱 빠르고 효율적으로 처리할 수 있다. 이와 같은 기술 혁신은 머신 러닝 및 딥 러닝과 같은 데이터 중심의 컴퓨팅 분야에서의 응용 가능성을 크게 넓힐 것으로 기대된다.

2. 배경

현대의 고성능 컴퓨팅 응용 프로그램은 강화된 데이터 처리 능력을 요구하며, 이를 지원하기 위해 DRAM 의 작동 원리와 PIM 의 개념에 대한 이해가 필수적이다. 심층 신경망(DNN), 합성곱 신경망(CNN), 순환 신경망(RNN) 등과 같은 응용 프로그램은 계산을 준비하기 위해 주 메모리로부터 높은 대역폭의 데이터 전송을 필요로 한다. 그러나, 포장 기술의 물리적 한계와 추가적인 물리적 핀의 비용 증가로 인해, 메모리의 오프칩 대역폭을 단순히 증가시키는 것은 경제적으로나 기술적으로 어려운 실정이다. 이로 인한 계산 속도의 제한과 높은 에너지 소비는 특히 머신 러닝과 같은 데이터 집약적 응용 프로그램에서 메모리 대역폭에 큰 부담을 일으킨다.

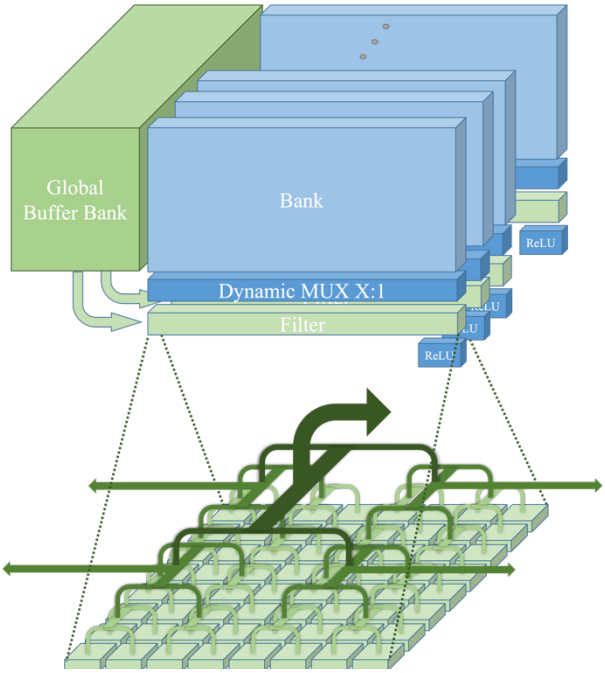
이러한 문제에 대응하여 메모리 내에서 직접 연산을 수행하여 메모리 대역폭을 효율적으로 활용하는 PIM 아키텍처에 대한 관심이 증가하고 있다. PIM 아키텍처는 DRAM 과 같은 전통적인 메모리 구조 내에 계산 기능을 통합하여, 데이터 전송 지연을 줄이고 시스템의 전체적인 에너지 효율을 향상시키는 것을 목표로 한다. SK 하이닉스가 개발한 Newton 은 이러

한 PIM 기술의 실제 적용 예로, 메모리 하드웨어 내에 곱셈 및 덧셈 연산기를 배치하여 행렬 연산을 타겟으로 가속화하는 Accelerator-in-Memory (AiM) 아키텍처를 제공한다.

본 연구의 아키텍처는 SK 하이닉스의 기존 아키텍처를 바탕으로 발전시켜, 메모리 내에서의 합성곱 신경망 병렬 데이터 처리 능력을 극대화하기 위한 목적으로 개발되었다. 기존의 구조에서 각 메모리 뱅크 내에서 독립적으로 연산을 수행하는 방식에 영감을 받아, 본 연구에서는 이러한 병렬성을 한층 더 확장하여 각 뱅크에서 여러 데이터 경로를 동시에 처리할 수 있는 분기(branch) 기능을 추가하였다. 이러한 분기 구조는 각 단계에서 데이터를 효율적으로 분할하여 처리함으로써, 복잡한 합성곱 연산에서 요구되는 대규모 데이터를 보다 빠르고 정확하게 처리할 수 있게 설계되었다.

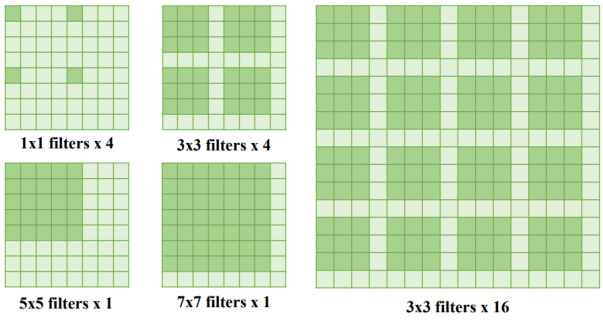
3. PINN 구조

PINN 구조는 특히 합성곱 신경망의 성능을 극대화하기 위해 설계되었다. 이 아키텍처는 기존의 DRAM을 활용하면서 메모리 내에서 데이터 처리를 수행함으로써 복잡한 합성곱 신경망 모델의 연산 속도를 혁신적으로 향상시키는 것을 목적으로 한다. 각 메모리 뱅크에 독립적으로 배치된 PINN 필터들은 각각의 데이터 스트림을 동시에 병렬로 처리할 수 있어, 전체적인 처리 효율을 크게 증가시킨다. 병렬 처리의 핵심 기능 중 하나는 Multiply-Accumulate(MAC) 연산 도중 데이터 스트림을 분기하여, 이를 통해 여러 데이터 처리 경로에서 다양한 필터 크기의 합성곱 연산을 수행할 수 있는 구조를 포함한다는 점이다. 이러한 구조는 각 분기점에서 독립적으로 연산을 진행함으로써 데이터 처리를 병렬화하고, 전체 연산 시간을 단축시키는 데 기여한다.



(그림 1) PINN 구조

또한, 이 아키텍처는 각 PINN 필터에 통합된 ReLU(비선형 활성화 함수) 유닛을 특징으로 한다. 이 유닛들은 MAC 연산 후 즉시 ReLU 연산을 적용함으로써, 다음 네트워크 계층으로 데이터를 전송하기 전에 필요한 비선형 변환을 수행한다. 이는 합성곱 신경망의 활성화 단계에서 중요한 역할을 하며, 전체 네트워크의 정확도와 반응 속도를 향상시키는 중요한 요소로 작용한다. 해당 디자인은 최신 네트워크의 활성화 단계가 연산 시간의 30%이상을 차지한다는 점에 착안하여, 활성화 단계에 특화된 연산 유닛을 배치함으로써 추가적인 성능향상을 도모하였다. PINN 아키텍처의 구조는 (그림 1)에 자세히 나타나 있으며, 이 그림은 PINN 아키텍처의 구성 요소들과 그들이 어떻게 상호 연결되어 있는지를 시각적으로 보여준다.

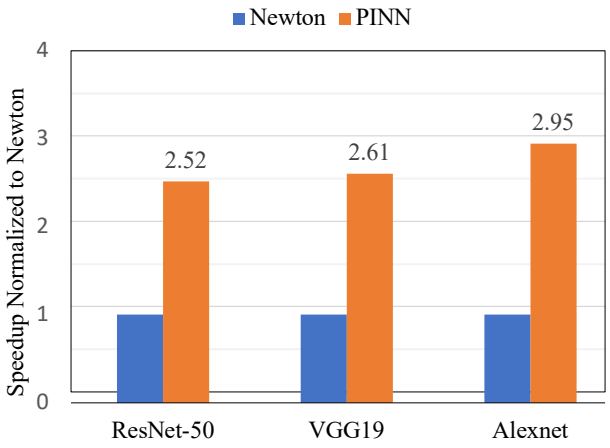


(그림 2) 연산 병렬화 방식

(그림 2)는 PINN 구조가 어떻게 CNN 합성곱 연산의 병렬화를 실현하는지를 보여준다. 이 구조는 각각의 블록을 효율적으로 활용하여, 동시에 여러 합성곱 필터 연산을 수행할 수 있도록 설계되었다. 왼쪽 기준으로, 1x1, 3x3 필터는 4 개를 한 CPU 사이클 내에 처리할 수 있는 능력을 갖추고 있어, PINN 아키텍처의 병렬 처리 능력을 극대화한다. 이런 구조는 뛰어난 확장성을 갖기에, PINN 필터를 확장한다면 오른쪽과 같이 3x3 필터는 최대 16 개를 한 번에 연산할 수 있는 구조를 어렵지 않게 만들 수 있다. 이러한 구조는 기존의 순차적 데이터 처리 방식에 비해, 데이터 처리량과 속도를 크게 향상시켜, 머신 러닝 작업의 효율을 크게 높인다. (그림 2)의 시각적 표현은 병렬 연산의 이점을 직관적으로 이해하는 데 도움을 준다.

4. 실험 결과

본 연구에서 수행된 실험 결과는 PINN 아키텍처가 합성곱 연산 처리에 있어 SK 하이닉스의 Newton 아키텍처와 어떻게 비교되는지를 보여준다. ResNet-50, VGG19, AlexNet 의 대표적인 합성곱 신경망 모델들의 연산 수를 기반으로 한 평가에서, PINN 은 각각의 네트워크에 필요한 합성곱 연산을 더 빠르게 수행할 수 있었다. 이러한 향상된 처리 능력은 PINN 이 각 메모리 뱅크에서 병렬로 수행할 수 있는 분기 기능을 활용함으로써 가능했으며, 이는 특히 CNN 연산의 병렬화와 효율성 측면에서 Newton 보다 우수한 성능을 나타냈다.



(그림 3) PINN 과 Newton 의 합성곱 신경망 성능 비교

(그림 3)은 ResNet-50, VGG19, AlexNet 모델에서 PINN 아키텍처와 Newton 아키텍처의 성능을 비교한 결과를 보여준다. 8x8 구조의 PINN 아키텍처는 1x1 과 3x3 필터에 대해서는 한 번에 네 개의 필터 연산을 수행한다. 5x5 와 7x7 필터는 한 번에 한 개의 필터 연산이 가능하다. 11x11 필터는 한 연산에 처리할 수 없으므로 두 번의 연산으로 처리한다. 이러한 구조는 Newton 대비 각각 2.52 배, 2.61 배, 2.95 배의 속도 향상을 VGG19, AlexNet, ResNet-50 모델에서 달성함을 보여준다. 이 수치는 PINN 아키텍처의 병렬 처리 능력이 Newton 대비 어떻게 딥러닝 연산 성능에 영향을 미치는지를 정량적으로 분석한 것이다.

5. 토론

본 연구의 결과는 PINN 아키텍처가 메모리 내에서 합성곱 신경망의 연산 속도를 혁신적으로 향상시켰음을 보여준다. 이 아키텍처는 데이터를 병렬로 처리하고, 합성곱 신경망에 최적화된 MAC 연산 구조를 통해 전통적인 시스템에서 발생하는 지연 시간을 상당히 줄이는 한편, CPU 와 메모리 간 데이터 전송을 최소화하여 전체적인 에너지 소비를 감소시키는 효과를 가져왔다. 그러나 PINN 기술의 구현은 기존의 Newton 아키텍처를 기반으로 하면서도, 분기 구조와 ReLU 연산기를 추가적으로 통합함으로써, 에너지 소비뿐만 아니라 여러 추가적인 기술적 과제를 야기할 수 있다. 이러한 구조의 복잡성 증가는 제조 비용의 상승과 수율 감소를 초래할 수 있으며, 이는 칩의 단가를 높이는 주요 요인이 된다.

복잡한 메모리 설계는 시스템 전체의 에너지 소비는 감소할 수 있으나, 메모리 칩 자체의 에너지 소비는 증가할 것으로 예상된다. 이러한 에너지 소비 증가는 메모리 칩의 발열 문제를 야기할 수 있으며, 고성능 컴퓨팅 환경에서는 이를 새로운 도전 요소로 고려해야 한다.

시스템의 복잡도가 증가함에 따라, 신호의 최장 지연 시간 또한 연장되는 결과를 낳는다. 이는 클록 속도의 증가를 제한하는 주요 요인으로 작용하여, 고속 클록을 요구하는 현대의 시스템 설계에서 전체 성능의 향상을 막는 장애물이 될 수 있다. 따라서, 설계자는 이러한 기술적 제약을 극복하고 PINN 아키텍처의 이점을 최대화하기 위한 방안을 모색해야 할 것이다.

6. 관련 연구

최근 몇 년간 PIM 기술에 대한 연구가 활발히 진행되어 왔다. 그중 'HBM-PIM' [2] 은 상용 DRAM 기술을 기반으로 한 PIM 의 하드웨어 아키텍처와 소프트웨어 스택에 대해 상세히 논의하였다. 이 연구에서는 PIM 기술을 실제 시스템에 통합하여 머신 러닝 애플리케이션과 그 외 다양한 애플리케이션의 성능을 향상시키는 방법을 제시한다.

또한, 본 연구의 참고가 되는 Newton [1] 에서는 AIM 아키텍처를 소개하면서, 이를 통해 가장 널리 사용되는 연산인 행렬 연산의 속도를 크게 향상시킬 수 있는 구조적 특징과 성능 개선 결과를 공유했다. Newton 은 기존 메모리 구조를 활용하여 현실적인 PIM 구조를 제시하는 것이 특징이다.

'SPACE' [3] 는 locality-aware 메모리 처리 기술을 통해 PIM 을 활용한 효율적인 데이터 처리 방법을 소개한다. 이 기술은 메모리 성질의 차이에 기반하여 자주 접근하는 데이터는 고성능 메모리에 저장하는 최적화를 한다. 본 연구는 개인화 추천시스템에 특화하여 성능향상을 보인다.

마지막으로, 'TensorDIMM' [4] 아키텍처에서는, PIM 을 활용하여 텐서 연산을 메모리에서 직접 처리할 수 있는 구조를 제시한다. 이 연구는 특히 딥 러닝의 텐서 연산을 메모리에서 일부 실행하여, 전체 시스템의 성능을 향상시키는 데 중점을 두고 있다.

이러한 다양한 연구들은 PIM 기술의 발전과 함께, 메모리와 연산의 통합이 다양한 컴퓨팅 환경에서 어떻게 활용될 수 있는지를 보여주며, 이는 향후 연구의 방향성에 중요한 영향을 미칠 것이다.

7. 결론

이 연구는 PINN 아키텍처가 합성곱 신경망 작업에 대한 성능과 효율성을 향상시킬 수 있음을 보여주었다. 처리 속도의 기존 연구 대비 최대 2.95 배 성능 향상은 메모리 아키텍처 내에 다양한 연산 기능을 통합하는 것의 잠재력을 강조한다. 기술적 경계를 계속 밀어붙이면서, PINN 은 더욱 강력하고 효율적인 컴퓨팅 시스템을 추구하는 우리의 노력에서 중요한 진전을 대표한다. 향후 연구는 기술을 정제하고, 다양한 컴퓨팅 환경에 적응시키며, 초기 평가 중에 확인된 기술적 장애물을 극복하는 데 집중할 것이다.

참고문헌

[1] He, Mingxuan, Choungki Song, Ilkon Kim, Chunseok Jeong, Seho Kim, Il Park, Mithuna Thottethodi, and T. N. Vijaykumar. "Newton: A DRAM-Maker's Accelerator-in-Memory (AIM) Architecture for Machine Learning." 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2020.

[2] S. Lee, S. Kang, J. Lee, H. Kim, E. Lee, S. Seo, H. Yoon, S. Lee, K. Lim, H. Shin, J. Kim, S. O, A. Iyer, D. Wang, K. Sohn, N. Kim, “Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology”, in 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)
[3] Kal, Hongju, Seokmin Lee, Gun Ko, and Won Woo Ro. “Space: Locality-Aware Processing in Heterogeneous Memory for Personalized Recommendations.” 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), 2021.
[4] Kwon, Youngeun, Yunjae Lee, and Minsoo Rhu. “Tensordimm.” Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, 2019.

---

* 이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-01361, 인공지능대학원지원(연세대학교); No. 2022-0-00050, 데이터 플로우 구조 기반 PIM 의 실행 및 프로그래밍 모델 개발; No. RS-2023-00277060, 개방형 엣지 AI 반도체 설계 및 SW 플랫폼 기술개발; No. RS-2024-00395134, 차세대 AI 반도체를 위한 DPU 중심의 데이터센터 아키텍처). 또한 이 논문은 삼성전자의 지원을 받아 수행된 연구임.