Sparse Tensor 가속기의 모델링에 관한 연구 동향

하회리[1], 백윤흥[2]
[1]서울대학교 전기정보공학부, 반도체공동연구소 박사과정
[2]서울대학교 전기정보공학부, 반도체공동연구소 교수
hwr0501@snu.ac.kr, ypaek@snu.ac.kr

A Study on Modeling of Sparse Tensor Accelerators

Whoi Ree, Ha[1], Yunheung, Paek[1]
[1]Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center (ISRC), Seoul National University

요 약

Sparse한 데이터가 딥러닝에 자주 사용됨에 따라 다양한 sparse 텐서 가속기들이 연구되고 있다. 하지만 이런 sparse 텐서 가속기들은 특수 하드웨어 모듈을 채용하고 있고, 다양한 구조로 되어 있다. 또한, 가속기들의 효율성이 데이터의 sparsity에 따라 달라지기 때문에 서로의 직접적인 비교도 힘들다. 따라서 이 문제들을 해결하기 위해, sparse 텐서 가속기들을 모델링하여 서로를 비교하려는 연구들이 존재하며, 이 논문에서는 이에 관한 연구 동향을 서술하였다.

1. 서론

Sparse 데이터는 다양한 딥러닝 모델을 통해 다양한 분야에서 널리 사용됩니다. 중요하지 않은 모델 가중치를 제거하는 pruning 기술은 불필요한 가중치들을 0으로 저장하여 데이터의 sparsity를 유발합니다. 또한, 응용 프로그램의 특성도 데이터의 sparsity를 유발할 수 있습니다. 예를 들어 자연어 처리에서는 각 단어가 처리되기 전에 임베딩되어 입력 벡터가 매우 sparse 해집니다.

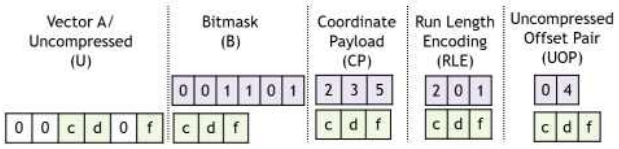
데이터의 sparsity는 텐서 계산에서 새로운 최적화 문제를 제기합니다. 입력 또는 가중치 중 하나가 0이면 다른 하나가 무엇이든 상관없이 출력이 0이 되므로 중복 계산을 유발합니다. 또한, 결과적으로 0은 최종 출력에 영향을 미치지 않습니다. 따라서 많은 연구가 이러한 sparsity를 이용하여 더 나은 효율성을 달성하기 위해 새로운 sparse 텐서 가속기를 설계하고 있습니다. 이들은 데이터의 sparsity를 이용하는 다양한 방법을 제안하고 있습니다. 예를 들어, 데이터를 인코딩하여 메모리 간의 데이터 흐름을 최소화하고, 데이터 전처리를 통해 0이 아닌 데이터만 감지하여 비효율적인 계산을 찾으려고 합니다. 또한, 텐서 계산을 처리하는 processing element (PE)의 배열에서 부하를 균형 있게 분산시켜 특정 PE가 0만 처리하지 않도록 합니다.

2. 문제점

이러한 sparse 텐서 가속기로 인해 보고된 속도 향상에도 불구하고, 특정 응용 프로그램에 대한 새로운 가속기를 설계할 때 큰 문제가 있습니다. Sparse 가속기들은 데이터에서 sparsity를 활용하여 가속기의 성능을 향상하는데, 이를 처리하는 담당 모듈이 필요합니다. 따라서, 가속기 설계에서 특수 하드웨어 모듈이 필요합니다. 이는 가속기 설계가 크고 및 비정규적인 설계 공간을 탐색해야 한다는 것을 뜻합니다. 또한, 많은 sparse 가속기들은 미리 정의된 시나리오를 전제로 합니다. EIE[1]는 딥러닝 모델이 'Deep Compression' 알고리즘을 사용하여 저장된다고 가정하고 있고, Sparten[2]은 SCNN[3]과 유사한 아키텍처를 가정합니다. 따라서 이러한 가속기들에 대한 보고된 효율성을 직접 비교하는 것은 적절하지 않습니다. 마지막으로, 많은 연구가 sparse 가속기 효율성이 데이터의 sparsity 정도에 따라 다르다는 것을 보여주고 있습니다. 데이터 안의 0의 개수와 0의 패턴은 각각의 sparse 가속기에 큰 영향을 미칩니다. 딥러닝 모델의 sparsity는 제어 가능할 수 있지만, 각 응용 프로그램의 입력 데이터는 통제할 수 없습니다. 따라서, 설계 시 어떤 sparse 가속기능이 특정 응용 프로그램에 효과적인지 결정하는 것은 현실적으로 불가능합니다.

3. Sparse Tensor 가속기의 모델링

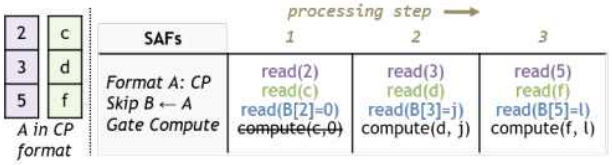
이러한 문제를 해결하기 위해서는 Sparse Tensor 가속기들을 모델링하여 최적의 설계를 탐색하는 것이 필요합니다. 특히 데이터의 sparsity를 사용하여 가속하는 sprase 가속 기능에 대한 모델링이 필요합니다. Sparse 가속 기능을 모델링하면 이에 따른 영향을 수학적으로 계산할 수 있고, 이를 통하여 현재 직접적으로 비교할 수 없는 다양한 가속 기능들을 서로 비교하여 어떤 것이 현재 상황에 맞는지 미리 계산할 수 있습니다.



<그림 1> Encoding Algorithms

Sparseloop[4]에서는 이런 문제를 해결하기 위해 몇 가지 sparse 가속 기능을 모델링 합니다. <그림1>에서는 sparseloop에서 지원하는 encoding 알고리즘들을 나타냅니다. Bitmask의 경우 0이 아닌 값들만 저장하고 metadata로 position의 위치를 같이 저장합니다. Coordinate payload도 마찬가지입니다. RLE의 경우 0이 아닌 값들 사이의 0의 개수를 metadata로 저장하며, UOP의 경우, 첫 번째와 마지막 0의 position을 metadata로 저장합니다. 이런 encoding 알고리즘을 sparse 텐서 가속기에서 많이 사용하는 이유는, 더 효율적으로 데이터를 나타내기 때문입니다. 실제로 Cambricon-x[5]와 같은 경우, encoding을 사용하여 DRAM access를 줄여 에너지와 performance적으로 더 효율적인 가속기를 설계하였습니다. 또한 0이 아닌 값과 metadata가 저장되기 때문에 이를 사용하여 더 효율적인 계산을 할 수 있습니다.

Sparseloop에서 지원하는 또 다른 가속 방식은 zero-skipping과 zero-gating입니다. Zero-gating은 0이 detection되면 PE를 실행하는 것이 아닌, idle상태로 두어 energy적인 면에서 더 효율적인 가속기를 만들 수 있습니다. Zero-skipping은 데이터의 값이 0인 경우, 다음 0이 아닌 값을 찾아와서 실행하는 방식입니다. 이 경우에는 energy와 performance 둘 다 효율적으로 만들 수 있지만, 0을 detection하고 다음 0이 아닌 값을 찾는 logic이 필요하기 때문에 가속기 설계가 더욱 복잡해집니다.



<그림 2> 모델링 예제

모델링 후 <그림 2>에서 볼 수 있듯이 각 가속 기능의 영향을 측정할 수 있습니다. 이 예제에서는 Coordinate payload 형식으로 저장된 데이터를 zero-gating과 zero-skipping이 적용된 상황을 보여주고 있습니다. 따라서 1번 step에서는 B[2]가 0이기 때문에 skipping되는 것을 보여주고 있고, 2,3 에서는 0 이 아니기 때문에 거기에 맞는 데이터를 불러와 실행하는 것을 모델링하고 있습니다. 이를 사용하면 총 latency나 energy consumption, utilization 등을 계산할 수 있습니다. 물론 완전히 정확하지는 않을 것이지만, 어느 정도 정확도로는 가속 기능을 평가 할 수 있게되고 이를 통해서 서로 다른 가속 기능에 대한 비교가 가능해집니다.

4. 결론

본 논문에서는 Sparse 텐서 가속기를 모델링해야 되는 당위성에 대해서 설명합니다. 딥러닝 모델과 딥러닝 모델에 사용되는 데이터는 sparse한 경우가 많기 때문에, sparse 텐서 가속기에 대한 연구가 활발히 진행되고 있습니다. 하지만 이런 sparse 텐서 가속기들은 데이터의 sparsity를 처리하기 위한 담당 모듈이 사용되고 있습니다. 따라서 크고 비정규적인 설계 공간을 가지고 있습니다. 또한, 데이터의 sparsity에 따라 가속 기능의 효율성이 달라지며, sparse 텐서 가속기들이 대부분 특정 시나리오를 가정하고 있기에 어떤 가속 기능이 효율적인지 비교할 수 없습니다. 따라서 가속 기능들을 모델링하여 영향을 계산할 수 있는 방식이 필요하며, 이를 위한 연구를 이 논문에서는 소개하였습니다.

ACKNOWLEDGEMENT

이 논문은 2023년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원, 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과 (IITP-2023-2020-0-01602)에 의하여 지원, 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.2020-0-01840,스마트폰의 내부데이터 접근 및 보호 기술 분석)

참고문헌

[1] Han, Song, et al. "EIE: Efficient inference engine on compressed deep neural network." ACM SIGARCH Computer Architecture News 44.3 (2016): 243-254.

[2] Gondimalla, Ashish, et al. "SparTen: A sparse tensor accelerator for convolutional neural networks." Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture. 2019.

[3] Parashar, Angshuman, et al. "SCNN: An accelerator for compressed-sparse convolutional neural networks." ACM SIGARCH computer architecture news 45.2 (2017): 27-40.

[4] Wu, Yannan Nellie, et al. "Sparseloop: An analytical, energy-focused design space exploration methodology for sparse tensor accelerators." 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). IEEE, 2021.

[5] Zhang, Shijin, et al. "Cambricon-X: An accelerator for sparse neural networks." 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2016.