Sparsity 를 활용한 DNN 가속기의 연구 동향

손선아[1], 강지은[1], 김소연[1], 김하늘[1], 김현정[1], 오현영[2*]
[1]가천대학교 AI 소프트웨어학부 학부생
[2]가천대학교 AI 소프트웨어학부 교수

{ wadadak99, kje147459, kimsoyeon744, znzl8722, qkqlguswjd , hyoh}@gachon.ac.kr

Recent Research in DNN Accelerators Exploiting Sparsity

Sun-Ah Son, Ji-Eun Kang, So-Yeon Kim, Ha-Neul Kim, Hyun-Jeong Kim, Hyunyoung Oh
Dept. of AI · Software, Gachon University

요 약

최근 딥러닝 연산의 고도화에 따라 희소성(Sparsity)을 효율적으로 처리할 수 있는 유연한 구조의 DNN 가속기가 중요해지고 있다. 그러나 기존의 가속기들은 유연성과 효율성 면에서 한계가 존재한다. 본 논문에서는 DNN 가속기의 기존 모델들과 최신 연구 동향에 대해 살펴본다. 특히 unstructured sparsity, structured sparsity, 그리고 최근 제안된 Hierarchical Structured Sparsity (HSS)를 적용한 가속기들을 분석하며, 각 접근 방식의 장단점을 비교한다.

1. 서론

효과적인 대규모 연산을 위해 DNN 가속기의 필요성이 대두되고 있다. 이는 딥러닝 연산에 특화된 구조로 연산 성능을 극대화하고 에너지 소모를 줄이는데 탁월하다. 이때 sparsity 를 효율적으로 활용할 수 있는 높은 유연성이 필요한데, 기존 가속기의 경우 종종 비효율적이거나 유연하지 못한 경우가 발생한다. 최근에는 이러한 한계를 극복하는 다양한 가속기에 대한 연구가 활발히 진행되고 있다. 본 논문에서는 structured 및 unstructured sparsity 를 활용한 기존 연구들과 함께, 새로운 sparsity 패턴인 HSS 를 적용한 HighLight 가속기를 소개한다.

2. 관련 연구

2.1 Unstructured Sparsity

Unstructured sparsity 는 activation function 이나 pruning 방법을 통해 입력 데이터에 따라 동적으로 발생한다. 따라서 DNN 모델 내에서 0 인 값의 위치가 불규칙하며 예측이 불가하다. 따라서, 효율적인 병렬처리가 어렵고, 하드웨어는 비정형적으로 분포된 non-zero 위치를 식별하고 처리해야 한다.

Unstructured sparsity 를 활용하기 위해 다양한 가속기 설계가 연구되고 있다. 대표적인 예로 SCNN [1], SIGMA [2], DSTC [3] 등이 있다.

SCNN 은 compressed sparse CNN 을 위한 가속기로, 입력 activation 과 weight 의 sparsity 를 모두 활용한다. 이 설계는 sparse matrix multiplication 을 위한 Cartesian product 기반의 접근 방식을 사용하여 불필요한 연산을 효과적으로 제거한다. SIGMA 는 DNN 학습을 위한 sparse and irregular GEMM (General Matrix Multiplication) 가속기이다. 이 설계는 다양한 sparsity 패턴을 지원하기 위해 유연한 와이어링을 사용하며, 런타임에 동적으로 변화하는 sparsity 를 처리할 수 있다. DSTC (Dual-side Sparse Tensor Core)는 outer product 기반의 연산 원리로 병렬연산을 지원하며, bitmap-based 인코딩 포맷을 결합하여 irregular 메모리 접근을 최소화한다. 기존의 tensor core 하드웨어에 최소한의 변경만으로 dual-sided sparsity 를 효과적으로 활용할 수 있으며, 행렬 곱 및 convolution 연산(SpCONV)에서 최대 10 배까지 성능을 향상시킨다.

Unstructured sparsity 는 activation sparsity 와 sparsity tax (overhead) 두 가지 측면에서 연산을 크게 줄일 수 있으나 cost 측면에서 문제가 발생한다. 또한 유연하다는 장점으로 다양한 sparsity 패턴을 처리하는 데 있어 우위를 점한다. 그러나 확장성과 효율성 측면에서 한계가 발생한다. 이러한 문제를 해결하려면 overhead 를 최소화하면서도 sparsity 를 효과적으로 활용해야 한다.

2.2 Structured Sparsity

Structured sparsity 는 DNN 모델의 가중치나 활성화 값에서 0 이 특정 규칙이나 공간적 제약에 따라

* 교신저자

분포됨을 말한다. 주로 pruning 의 skip 의 기법을 사용해 pruning 된 요소(0 의 값)를 건너뛰며 연산의 효율성을 높인다.

Structured sparsity 를 활용한 대표적인 가속기 연구로는 NVIDIA 의 Sparse Tensor Core (STC) [4], S2TA [5], Sparse Tensor Core(Zhu et al.)[6] 등이 있다. NVIDIA 의 STC 는 2:4 sparsity 패턴을 이용해 4 개의 값 중 2 개만 0 을, 나머지 2 개는 0 이 아님을 이루어 50%의 sparsity 를 갖는다. 이 방식은 간단하면서도 효과적인 하드웨어 구현이 가능하여 낮은 sparsity tax 를 달성한다. S2TA (Structured Sparsity Tensor Accelerator)는 구조화된 sparsity 를 활용하여 모바일 CNN 가속을 위한 설계다. 이 가속기는 weight 와 activation 모두에 구조화된 sparsity 를 적용하여 에너지 효율성을 높인다. Sparse Tensor Core(Zhu et al.)는 벡터 단위 sparse 신경망을 위한 알고리즘과 하드웨어 co-design 방식을 제안한다. 이 접근법은 현대 GPU 에서 sparse 텐서 연산의 성능을 향상시키는 것을 목표로 한다.

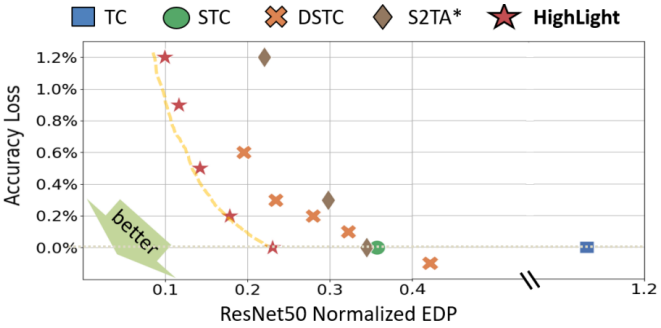
Structured sparsity 가속기의 장점은 0 이 아닌 값들을 규칙적으로 배치해 하드웨어가 쉽게 식별하고, 계산 유닛에 효율적으로 분배하는 것이다. 이를 통해 하드웨어 부하의 감소와 최적화된 메모리 사용량을 달성한다. 그러나 한정된 특정 sparsity 패턴만 처리할 수 있으며, 고정된 형태로 pruning 을 적용하기 때문에 unstructured sparsity 에 비해 유연성이 다소 떨어진다는 한계점을 갖는다.

2.3 HighLight (Hierarchical Structured Sparsity)

최근 structured 구조이면서 이전 모델과는 다르게 유연성과 효율성을 모두 갖춘 모델에 대한 논문이 발표되었다[7]. 이 논문에서는 새로운 종류의 sparsity 인 HSS(Hierarchical Structured Sparsity)를 도입한 HighLight 가속기를 제안한다. HSS 와 같은 간단한 sparsity 패턴은 가속 기능을 구현하는 하드웨어를 단순화하여 낮은 sparsity tax 를 유지하는 데 도움이 된다. HSS 의 경우 파이버트리(fibertree) 추상화 기반의 계층적 트리 구조를 사용하여 하나의 랭크에 서로 다른 sparsity 패턴을 적용할 수 있다.

HSS 기반 희소화의 목표는 0 이 아닌 값을 가능한 많이 보존하는 것이다. 목표를 달성하기 위해 고밀도 텐서를 낮은 순위에서 높은 순위로 순위별로 희소화한다. 예를 들어, C1 (3:4)→C0 (2:4) HSS 의 경우, 먼저 C0 의 2:4 패턴을 적용한 다음 C1 의 3:4 패턴을 적용한다. 이러한 방식으로 순위별 sparsity 패턴에 따른 HSS 의 유연성으로 다양한 sparsity 정도를 가진 sparse 모델을 얻을 수 있다.

HighLight 는 다양한 sparsity degree 에서 기존 설계들보다 더 나은 성능을 보인다. 그림 1 에서 볼 수 있듯이, HighLight 는 ResNet-50 모델에 대해 EDP(energy-delay product)-정확도 손실 Pareto frontier 상에 위치한다. [7] 논문의 다른 비교에서 STC 는 50% sparsity 에서만 좋은 성능을 보이고, DSTC 는 유연성은 높지만 밀집된 모델에서는 비효율적이다. S2TA 는 구조화된 sparsity 를 활용하지만 순수 dense 레이어 처리에 제한이 있다. HighLight 는 다양한 sparsity degree 에서 6.4 배(최대 20.4 배)의 더 나은 EDP 를 달성하며, sparse 가속기와 비교해 2.7 배(최대 5.9 배)의 EDP 개선을 보인다.



(그림 1) Highlight 의 EDP-정확도 성능 비교[7]

3. 결론

본 논문에서는 unstructured 와 structured sparsity 기반의 가속기 논문들을 살펴보았다. 기존 여러 가속기는 효율성과 유연성을 동시에 갖추기 어려웠으나 HighLight 논문에서는 계층적 트리 구조를 사용하여 한계점을 효과적으로 극복하여 균형을 달성하였다. 해당 논문에서는 하드웨어 지원에 대해 논의를 하지 않았는데, 이러한 부분에 대한 연구가 더 추가된다면 DNN 가속기의 성능 발전에 기여할 수 있을 것으로 기대된다.

사사문구

이 논문은 2024 년도 정부(산업통상자원부)의 재원으로 한국산업기술기획평가원의 지원(No. RS-2024-00406121, 자동차보안취약점기반위협분석시스템개발(R&D))과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. RS-2022-00166529)을 받고 과기정통부 정보통신기획평가원의 정보보호핵심원천기술개발사업(No. RS-2024-00337414)으로 수행한 결과임.

참고문헌

[1] A. Parashar et al., "SCNN: An Accelerator for Compressed-Sparse Convolutional Neural Networks," ISCA, 2017

[2] E. Qin et al., "SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training," HPCA, 2020

[3] Y. Wang et al., "Dual-side Sparse Tensor Core," ISCA, 2021

[4] NVIDIA, “NVIDIA Ampere GA102 GPU Architecture,” Technical Report, 2020.

[5] Liu et al., “S2TA: Exploiting Structured Sparsity for Energy-Efficient Mobile CNN Acceleration,” HPCA, 2022

[6] Maohua Zhu et al., “Sparse Tensor Core: Algorithm and Hardware Co-Design for Vector-wise Sparse Neural Networks on Modern GPUs,” MICRO, 2019

[7] Yannan Nellie Wu, Et Al, “HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity,” MICRO, 2023