엣지 디바이스를 위한 AI 가속기 설계 방법

하회리 [1], 김현준 [1], 백윤흥 [2]
[1] 서울대학교 전기정보공학부, 반도체공동연구소 박사과정
[2] 서울대학교 전기정보공학부, 반도체공동연구소 교수

wrha@sor.snu.ac.kr, hjkim@sor.snu.ac.kr , ypaek@snu.ac.kr

AI Accelerator Design for Edge Devices

Whoi Ree, Ha[1], Hyunjun Kim[1], Yunheung Paek[1]
[1]Dept. of electrical and Computer Engineering and Inter-university Semiconductor Research Center, Seoul National University

요 약

단일 dataflow 를 지원하는 DNN 가속기는 자원 효율적인 성능을 보이지만, 여러 DNN 모델에 대해서 가속 효과가 제한적입니다. 반면에 모든 dataflow 를 지원하여 매 레이어마다 최적의 dataflow 를 사용하여 가속하는 reconfigurable dataflow accelerator (RDA)는 굉장한 가속 효과를 보이지만 여러 dataflow 를 지원하는 과정에서 필요한 추가 하드웨어로 인하여 효율적이지 못합니다. 따라서 본 연구는 제한된 dataflow 만을 지원하여 추가 하드웨어 요구사항을 감소시키고, 중복되는 하드웨어의 재사용을 통해 최적화하는 새로운 가속기 설계를 제안합니다. 이 방식은 자원적 한계가 뚜렷한 엣지 디바이스에 RDA 방식을 적용하는데 필수적이며, 기존 RDA 의 단점을 최소화하여 성능과 자원 효율성의 최적점을 달성합니다. 실험 결과, 제안된 가속기는 기존 RDA 대비 32% 더 높은 에너지 효율을 보이며, latency 는 불과 1%의 차이를 보였습니다.

1. 서론

딥 뉴럴 네트워크(DNN)는 콘텐츠 추천[1], 얼굴 인식[2], 챗봇[3]을 포함한 많은 AI 기반 응용 프로그램에서 필수적인 도구로 자리잡았습니다. DNN 알고리즘이 지속적으로 발전함에 따라, DNN 모델의 규모와 복잡도가 증가하고 있고, 더 많은 계산 능력을 필요로 합니다. 특히 이 문제는 자원적 한계가 뚜렷한 엣지 디바이스에서 더 중요해집니다. 또한 DNN 모델이 다양한 방식으로 발전함에 따라, 각각의 DNN 모델은 고유한 계산 특성을 보여주고 있습니다[4], [5]. 따라서, 다양한 AI 기반 애플리케이션을 지원하기 위해 제한된 자원 내에서 여러 DNN 을 효율적으로 계산할 수 있는 가속기의 개발이 필요합니다.

DNN 의 dataflow 는 각 피연산자에 대해 세 가지 유형인 input-stationary, weight-stationary, output-stationary 로 분류될 수 있습니다. 효율성을 극대화하기 위해, DNN 계산은 각 dataflow 유형에 따른 해당 차원에서 병렬 처리될 수 있습니다. 예를 들어, Shi-diannao[6]는 output 채널 차원을 통한 병렬 처리를 활용하여 partial output 의 읽기 및 쓰기를 최소화하는 output-stationary dataflow 을 사용합니다. 따라서, Shi-diannao 는 이 dataflow 을 지원하기 위해 각 PE 내에 output 전용 레지스터와 덧셈기를 가지고 있습니다. 이렇게 하나의 dataflow 을 선택하고 지원하는 가속기를 fixed dataflow accelerator (FDA)라고 합니다.

그러나 FDA 의 단점은 단일 dataflow 가 다양한 유형의 레이어와 모델에 대해 최적의 성능을 보장하지 못한다는 것입니다 [7], [8]. PE 의 수, 캐시 크기 및 메모리 대역폭과 같은 하드웨어 변수들과 input, weight, output 의 형태와 같은 소프트웨어 변수들이 특정 dataflow 의 효율성을 결정하는 데 기여합니다. DNN 은 다양한 유형 및 크기의 레이어를 가지고 있기 때문에, FDA 의 효율성은 그에 따라 변동됩니다. 예를 들어, 모델의 input 및 output 채널 차원이 PE 배열 크기의 배수가 아닌 경우, Shi-diannao 의 utilization 이 감소합니다[6].

이 문제를 해결하기 위해 다양한 모델의 요구 사항에 맞게 dataflow 를 동적으로 조정할 수 있는 reconfigurable dataflow accelerator (RDA)에 대한 연구가 진행되었습니다. 예를 들어, Eyeriss v2[9]는 현재 레

이어와 데이터 유형의 형태에 기반하여 dataflow 를 동적으로 조정할 수 있습니다. 결과적으로, Eyeriss v2 는 고정된 dataflow 을 지원하는 이전 버전에 비해 12.6 배 더 높은 throughput 을 달성하였습니다. 그러나 가속기를 reconfigure 하여 각 레이어에 대한 dataflow 를 최적화하는 것은 상당한 에너지 오버헤드를 초래합니다[10]. 또한, reconfiguration 을 위해 필요한 추가 하드웨어 구성 요소로 인해 RDA 의 칩 면적은 FDA 보다 훨씬 커질 수밖에 없습니다. 따라서 RDA 는 성능면에서는 FDA 보다 월등하지만, 자원 효율적 관점에서는 취약하기 때문에 엣지 디바이스에서는 사용하기 어렵습니다.

이 연구에서는 RDA 의 단점을 최소화하여 성능과 자원 효율적 관점에서 최적의 design 을 찾습니다. 기존 RDA 와 같이 모든 dataflow 를 지원하는 것이 아닌, 제한된 dataflow 만 지원하여 기존의 비효율성을 최소화합니다. 제한된 dataflow 만 지원하기 때문에, 필요한 추가적 하드웨어가 줄어들고, 동적으로 reconfigure 하는 과정에서 발생하는 성능적 overhead 도 최소화할 수 있습니다. 또한, 추가적 하드웨어를 분석하여 중복되는 하드웨어를 찾아 재사용하는 방식의 최적화 기법을 적용합니다. 실험 결과, RDA 대비 32% 더 높은 에너지 효율을 보이지만 latency 는 1%밖에 차이 나지 않는 가속기를 설계하였습니다.

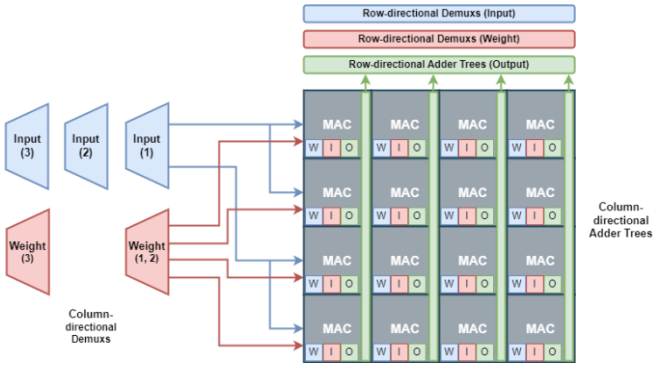
2. 배경 지식

Dataflow 는 버퍼 접근 횟수, 계산의 병렬성, 데이터 재사용 등을 결정함으로써 가속기의 전반적인 효율성을 결정합니다 [9], [11]. Dataflow 는 temporal 과 spatial 매핑으로 구성됩니다. Temporal 매핑은 데이터의 순서 및 타일링을 나타냅니다. 메모리 크기와 대역폭에 따라 타일링은 데이터 fetching 을 결정하고, 순서는 루프 계산의 배열을 결정합니다. 반면, spatial 매핑은 PE 배열에 제공되는 데이터의 크기와 차원을 나타냅니다. 이는 몇 개의 계산이 동시에 실행되는지를 정의함으로써 가속기의 utilization 에 직접적인 영향을 미칩니다.

최적의 dataflow 는 레이어의 형태, 작업, 크기에 따라 달라지기 때문에, 하나의 고정된 dataflow 는 여러 레이어의 DL 모델, 그리고 여러 모델에 대해 이상적일 수는 없습니다 [12], [11], [8]. 따라서, 최신 RDA 인 MAERI [10]는 모든 dataflow 을 지원할 수 있도록 프로그래밍 가능한 완전히 dataflow-flexible 한 가속기를 설계하였습니다. MAERI 는 reconfigurable 한 덧셈 트리 네트워크와 스위치 및 유연한 NoC 모듈을 사용하여 DNN 에서 사용 가능한 모든 dataflow 를 지원할 수 있습니다. 이 architecture 를 사용하여 레이어별로 최적의 dataflow 매핑을 변경함으로써, MAERI 는 최신 FDA 에 비해 8~459% 더 나은 utilization 을 달성하였습니다.

그러나 저자들이 후속 연구에서 지적한 바와 같이 [13], reconfigurable 한 architecture 는 overhead 를 유발합니다. 스위치와 전선과 같은 추가 하드웨어 구성 요소가 필요하며, 이는 상당한 에너지 및 면적 오버헤드를 발생시킵니다. 실제로, MAERI 의 아키텍처는 매우 비효율적입니다. 모든 dataflow 를 지원하기 위해, MAERI 는 아키텍처 내 모든 곱셈기와 덧셈기에 스위치를 추가합니다. 또한, 최적의 dataflow 에 따라 input/weight 를 적절한 곱셈기에 공급하기 위해, 그 곱셈기 수의 두 배에 해당하는 스위치 트리가 필요합니다. 보고된 바에 따르면, MAERI 는 NVDLA 스타일의 FDA 에 비해 평균적으로 11.7% 더 많은 에너지를 필요로 합니다. 또한, 레이어별로 reconfigure 하는 것은 각 레이어 실행의 끝에 추가적인 지연 및 전력 오버헤드를 나타냅니다.

3. 디자인



(그림 1) Architecture Design

<그림 1>은 엣지 디바이스를 위한 AI 가속기 디자인을 보여줍니다. 각 PE 는 MAC 유닛과 weight, input, output 을 위한 버퍼로 구성됩니다. 행 방향 및 열 방향의 디멀티플렉서와 덧셈 트리는 지정된 dataflow 에 맞춰 적절한 데이터를 제공합니다. 가속기에서 지원하는 dataflow 는 정적 분석을 통해 결정되기 때문에, 적절한 PE 에 디멀티플렉서와 덧셈 트리를 배선함으로써 dataflow 에 맞춰 데이터를 계산할 수 있습니다. 간단히 말해서, 우리가 지원하는 각 dataflow 는 이에 맞는 전용 디멀티플렉서 와 덧셈 트리 세트를 가지고 있습니다. 그런 다음 적절한 디멀티플렉서 및 덧셈 트리 세트를 선택함으로써 dataflow 를 동적으로 전환할 수 있습니다. 이는 각 디멀티플렉서 및 덧셈기에 활성화/비활성화 신호를 보내는 것으로 수행됩니다.

엣지 디바이스에서 또 다른 중요한 기준은 칩 면적입니다. 다양한 dataflow 을 동적으로 지원하기 위해서

는 추가적인 하드웨어 구성 요소가 필요하기 때문에, 우리는 가속기가 지원하는 dataflow 의 수를 제한합니다. 또한, dataflow 전환에 관여하는 하드웨어 구성 요소의 재사용을 최대화함으로써 가속기를 더욱 최적화합니다. 이러한 접근 방식은 하드웨어 자원의 효율적 사용을 가능하게 하여, 제한된 칩 면적 내에서 최대한의 성능을 발휘할 수 있도록 합니다. Dataflow 간 전환에 사용되는 구성 요소들의 재사용은 전체적인 하드웨어 비용을 줄이는 동시에, 다양한 dataflow 을 효과적으로 처리할 수 있는 유연성을 제공합니다. 따라서, 칩 면적과 성능 사이의 균형을 최적화하는 것이 중요하며, 이는 엣지 디바이스의 DNN 가속기 설계에서 핵심적인 고려 사항 중 하나입니다.

어떤 경우에는 dataflow 들이 동일한 데이터 fetching 패턴을 공유할 수 있습니다. 이는 dataflow 간의 하드웨어 구성 요소의 중복이 발생한다는 것을 뜻하고, 이를 줄여 전체 칩 면적을 최소화할 수 있습니다. 예를 들어, <그림 1>에서 볼 수 있듯, dataflow 1 과 dataflow 2 는 같은 열 방향 weight 매핑을 가지고 있습니다. 이 경우에는 별도의 디멀티플렉서를 할당하지 않고, 같은 디멀티플렉서를 재사용합니다. 재사용 가능한 디멀티플렉서를 식별하는 것은 dataflow 의 spatial 매핑을 살펴봄으로써 이루어질 수 있습니다. 마찬가지로, 행 방향 디멀티플렉서와 덧셈 트리에 대해서도 중복된 하드웨어를 탐색하여 공유할 수 있게 design 하여 area 및 energy overhead 를 최소화하였습니다.

4. 실험

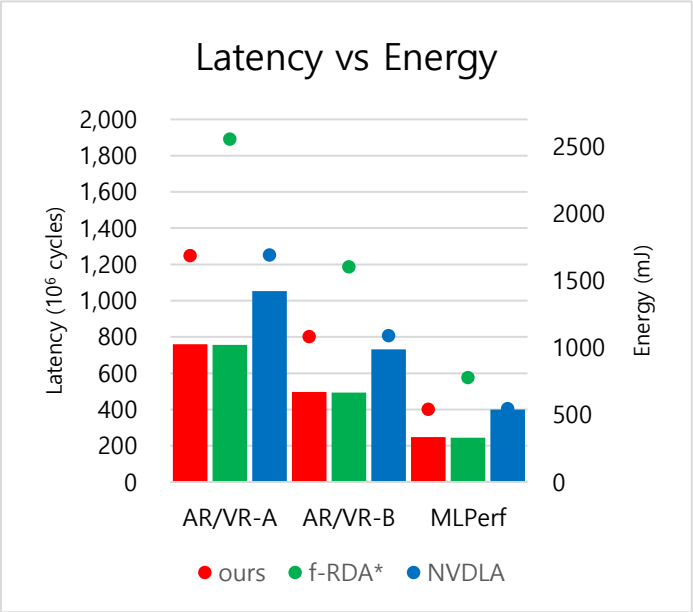
Zigzag [14]라는 state-of-the-art design space exploration (DSE) 툴을 사용하여 dataflow 들을 정적 분석하고 어떤 dataflow 들을 지원할지 결정하였습니다. Zigzag 는 dataflow 의 utilization, energy consumption 등을 예측하여 각 dataflow 의 효율을 미리 계산할 수 있게 해줍니다. 또한 CACTI7[15]를 사용하여 메모리 유닛의 에너지와 크기를 계산하였고, 이 때 45nm technology 를 사용하였습니다. 또한 다른 하드웨어 요소들은 Synopsys 사의 Design Compiler 를 통하여 계산하였습니다.

하드웨어는 자원은 [13]에서 사용하는 엣지 디바이스에 맞게 32 x 32 차원의 PE array 와 16GB/s bandwidth, 4 MiB 의 on-chip 메모리를 가정하였습니다. 데이터는 8-bit 정수 형태를 가집니다. 또한 비교대상은 RDA 형식으로 구현한 f-RDA 와 FDA 인 NVDLA 가속기입니다.

<표 1> Workloads For Evaluation

Workload	Models
AR/VR-A	Resnet50 Unet MobileNetV2
AR/VR-B	Resnet50 Unet MobileNetV2 BR-Q DepthNet
MLPerf	Resnet50 MobileNetV1 SSD-Resnet34 SSD-MobileNetV1 GNMT

실험에 사용한 벤치마크는 <표 1>에 표시되어 있습니다. 기본적으로 현재 엣지 디바이스에서 사용하는 CNN 기반 DNN 으로 이루어져 있으며, 이를 AR/VR-A 와 AR/VR-B 로 구분하여 실험하였습니다. 또한 더 다양한 DNN 을 위하여 MLPerf 에서는 CNN 포함 GNMT 와 같은 CNN 이 아닌 다른 기반의 DNN 도 포함하여 실험하였습니다. 해당 벤치마크는 마찬가지로 [13]에서 인용하였습니다.



(그림 2) Latency vs Energy Comparison

<그림 2>는 latency 와 energy 를 비교한 결과 값입니다. 바 그래프는 latency 를 dot 은 energy 를 나타냅니다. Latency 와 energy 모두 낮은 값이 더 높은 성능을 뜻합니다. 결과적으로 이 연구에서 제안한 방식은 RDA 와 비교하였을 때, latency 는 0.95% 더 높지만, 32.27% 더 좋은 energy 효율을 보입니다. 또한 NVLDA 와 비교하였을 때, energy 는 0.52% 더 높지만, latency 를 31.67% 더 낮추었습니다. 결론적으로 latency 는 RDA 에 가깝지만 energy 효율은 FDA 인 NVDLA 와 가까운 효율적인 가속기를 설계하였습니다.

5. 결론

이 연구는 다양한 모델 요구사항에 맞추어 dataflow 를 동적으로 조정할 수 있는 reconfigurable dataflow accelerator (RDA)의 개선을 목표로 합니다. RDA 는 레이어별로 최적의 dataflow 로 전환하여 실행할 수 있는 가속기이기 때문에, 굉장한 가속 성능을 보입니다. 하지만 전환 과정에서 발생하는 에너지 오버헤드와 여러 dataflow 를 지원하기 위한 추가적인 하드웨어로 인해 칩 면적 증가가 문제로 지적되었습니다. 특히 이는 자원이 한정된 엣지 디바이스에 큰 문제가 됩니다. 이에 대응하여, 본 연구에서는 제한된 dataflow 만을 지원함으로써 하드웨어 요구사항을 감소시키고, reconfiguration 과정에서의 성능적 오버헤드를 최소화하는 동시에 하드웨어 재사용을 통한 최적화 기법을 적용하여 자원 효율성을 개선한 새로운 가속기를 제안하였습니다. 실험 결과, 이 가속기는 기존 RDA 대비 32% 더 높은 에너지 효율을 달성하면서 latency 는 거의 차이 나지 않는 성능을 보였습니다.

6. 사사문구

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2023-00277326), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (No.RS-2023-00277060, 개방형 엣지 AI 반도체 설계 및 SW 플랫폼 기술개발), 2024 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원, 반도체 공동연구소 지원, 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2023-RS-2023-00256081), 반도체 공동연구소 지원을 받아 수행된 연구임.

참고문헌

[1] Rappaz, Jérémie, Julian McAuley, and Karl Aberer. "Recommendation on live-streaming platforms: Dynamic availability and repeat consumption." *Proceedings of the 15th ACM Conference on Recommender Systems*. 2021.

[2] Balaban, Stephen. "Deep learning and face recognition: the state of the art." *Biometric and surveillance technology for human and activity identification XII* 9457 (2015): 68-75.

[3] Sperlí, Giancarlo. "A cultural heritage framework using a Deep Learning based Chatbot for supporting tourist journey." *Expert Systems with Applications* 183 (2021): 115277.

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

[5] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[6] Du, Zidong, et al. "ShiDianNao: Shifting vision processing closer to the sensor." *Proceedings of the 42nd annual international symposium on computer architecture*. 2015.

[7] Yang, Xuan, et al. "Interstellar: Using halide's scheduling language to analyze dnn accelerators." *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 2020.

[8] Parashar, Angshuman, et al. "Timeloop: A systematic approach to dnn accelerator evaluation." *2019 IEEE international symposium on performance analysis of systems and software (ISPASS)*. IEEE, 2019.

[9] Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." *IEEE journal of solid-state circuits* 52.1 (2016): 127-138.

[10] Kwon, Hyoukjun, Ananda Samajdar, and Tushar Krishna. "MAERI: Enabling Flexible Dataflow Mapping over DNN Accelerators via Reconfigurable Interconnects." *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*. 2018.

[11] Dave, Shail, et al. "Dmazerunner: Executing perfectly nested loops on dataflow accelerators." *ACM Transactions on Embedded Computing Systems (TECS)* 18.5s (2019): 1-27.

[12] Kwon, Hyoukjun, et al. "Understanding reuse, performance, and hardware cost of dnn dataflow: A data-centric approach." *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. 2019.

[13] Kwon, Hyoukjun, et al. "Heterogeneous dataflow accelerators for multi-DNN workloads." *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021.

[14] Mei, Linyan, et al. "ZigZag: Enlarging joint architecture-mapping design space exploration for DNN accelerators." *IEEE Transactions on Computers* 70.8 (2021): 1160-1174.

[15] Balasubramonian, Rajeev, et al. "CACTI 7: New tools for interconnect exploration in innovative off-chip memories." *ACM Transactions on Architecture and Code Optimization (TACO)* 14.2 (2017): 1-25.