Smart Media Journal / Vol.13, No.3 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2024.13.3.36

국내외 인공지능 반도체에 대한 연구 동향

(Research Trends in Domestic and International AI chips)

김현지*, 윤세영**, 서화정***

(Hyun Ji Kim*, Se Young Yoon**, Hwa Jeong Seo***)

요 약

최근 ChatGPT와 같은 초거대 인공지능 기술이 발달하고 있으며, 다양한 산업 분야 전반에서 인공지능이 활용됨에 따라 인공지능 반도체에 대한 관심이 집중되고 있다. 인공지능 반도체는 인공지능 알고리즘을 위한 연산을위해 설계된 칩을 의미하며, NVIDIA, Tesla, ETRI 등과 같이 국내외 여러 기업에서 인공지능 반도체를 개발 중에 있다. 본 논문에서는 국내외 인공지능 반도체 9종에 대한 연구 동향을 파악한다. 현재 대부분의 인공지능 반도체는 연산 성능을 향상시키기 위한 시도들이 많이 진행되었으며, 특정 목적을 위한 반도체들 또한 설계되고 있다. 다양한 인공지능 반도체들에 대한 비교를 위해 연산 단위, 연산속도, 전력, 에너지 효율성 등의 측면에서 각 반도체에 대해 분석하고, 현재 존재하는 인공지능 연산을 위한 최적화 방법론에 대해 분석한다. 이를 기반으로 향후 인공지능 반도체의 연구 방향에 대해 제시한다.

■ 중심어 : 인공지능 ; 인공지능 반도체 ; 신경처리장치

Abstract

Recently, large-scale artificial intelligence (AI) such as ChatGPT have been developed, and as AI is used across various industrial fields, attention is focused on AI chips (semiconductors). AI chips refer to chips designed for calculations for AI algorithms, and many companies at domestic and abroad, such as NVIDIA, Tesla, and ETRI, are developing AI chips. In this paper, we survey research trends on nine types of AI chips. Currently, many attempts have been made to improve the computational performance of most AI chips, and semiconductors for specific purposes are also being designed. In order to compare various AI semiconductors, each chip is analyzed in terms of operation unit, speed, power, and energy efficiency. We introduce currently existing optimization methodologies for AI computation. Based on this, future research directions for AI semiconductors are presented in this paper.

■ keywords : Artificial Intelligence ; Artificial Intelligence Semiconductors, Neural Processing Unit (NPU)

Ⅰ. 서 론

최근 인공지능 기술이 급격히 발전함으로써 다양한 산업 분야 전반에서 인공지능 (AI)이 활용됨에 따라 인공지능 반도체 (AI Semiconductors, AI chips)에 대한 필요성이 또한 대두되고 있다. 인공지능 반도체는 인공지능 알고리즘을 위한 연산을 수행하기 위해 설계된 칩을 의미한다. 따라서 대부분의 인공지능 반도체는 딥러닝 작업의 효율성 및 가속화에 초점을 두고 있기 때문에 이외의 연산에는 사용되기 어렵다. 현재, 빠른 연산 속도를 달성하기 위한 연구들이 진행되고 있으며, 여러 종류가 있어 특정 딥러

* 학생회원, 한성대학교 정보컴퓨터공학과
** 학생회원, 한성대학교 융합보안학과
*** 정회원, 한성대학교 융합보안학과
This research was financially supported by Hansung University.

접수일자 : 2024년 02월 22일
수정일자 : 2024년 03월 08일
게재확정일 : 2024년 03월 20일
교신저자 : 서화정 e-mail : hwajeong84@gmail.com

닝 작업에 적합한 반도체를 사용해야한다.

본 논문에서는 인공지능 반도체의 특징에 대해 살펴보고, 국내외 인공지능 반도체에 대한 연구 동향 및 분석 그리고 향후 연구 방향을 제시한다.

Ⅱ. 본 론

1. 인공지능 반도체

인공지능 반도체에는 크게 4종류가 있으며, 그림 1과 같이 한 번에 연산할 수 있는 크기가 다르다. 본 절에서는 인공지능 연산을 위한 프로세서들의 특징에 대해 설명한다.

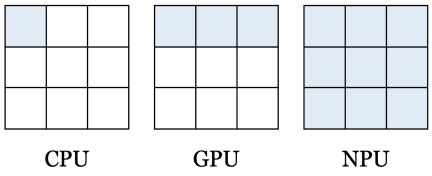

그림 1 각 프로세서 별 단일 연산 단위

가. GPU (Graphics Processing Unit) [1]

GPU는 복잡한 그래픽 계산을 처리하기 위해 설계되었지만, 벡터 단위의 연산을 병렬적으로 수행할 수 있다는 특징으로 인해 대량의 데이터를 처리하는 딥러닝에도 유용하게 사용되어 오고 있다.

나. 텐서코어 (Tensor Cores) [2]

NVIDIA에서 개발한 텐서 코어는 복잡한 행렬 연산을 빠르게 처리할 수 있도록 설계되었으며, 딥러닝 모델의 학습과 추론 작업에 필수적인 행렬 곱 연산을 가속화한다.

다. TPU (Tensor Processing Unit) [3]

Google이 개발한 TPU는 텐서 연산에 최적화된 인공지능 전용 반도체이다. 특히 구글의 딥러닝 프레임워크인 TensorFlow와 함께 사용될 수 있다. 이를 위한 경량화 모델인 TensorFlow Lite 모델을 사용할 때 높은 효율성을 보인다.

라. NPU (Neural Processing Unit) [4]

NPU는 인공지능에 필요한 연산을 특화하여 처리하는 하드웨어를 의미한다. 이는 딥러닝 모델에 필요한 연산을 가속화하고, 전체 시스템의 에너지 효율성을 높이는 데 중점을 둔다. 예를 들면 딥러닝 추론을 위한 NPU가 있으며, 이러한 칩은 해당 작업에서는 매우 효율적이지만, 다른 용도로는 사용하기 어려울 수 있다. NPU의 일반적인 구조는 그림 2와 같다. 연산을 위한 코어가 존재하며, 빠른 데이터 접근을 위한 On-Chip 메모리, 데이터 입출력을 위한 인터페이스 등으로 이루어져있다. 이외에도 연산 중간 값을 저장하는 버퍼가 존재하거나 세부 스펙이 다르게 설계될 수 있지만, 아래 그림과 같이 빠르게 데이터에 접근하기 위해 인터페이스를 거쳐 On-Chip 메모리에 데이터를 가져오고, 병렬 처리에 특화된 처리 코어는 공통적인 부분이다. 각 NPU 마다의 아키텍처의 구조에 따라 성능 차이가 나거나 적합한 작업이 달라질 수 있다. 본 논문에서 언급되는 제품들은 NVIDIA의 H100 (텐서코어 유형) 을 제외하면 모두 이 유형에 속한다.

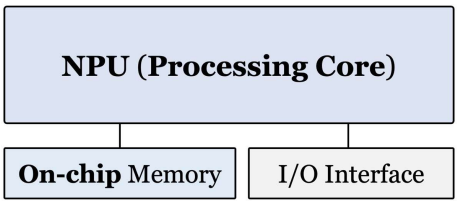

그림 2 일반적인 NPU의 공통적 구조

2. 인공지능 반도체의 주요 특징

인공지능 반도체에서 가장 중요하게 살펴볼 요소는 연산 단위, 에너지 효율성, 처리 속도이다. 이러한 요소들은 인공지능 기반 시스템의 성능 및 운영 비용에 직접적인 영향을 미치기 때문에 특정 작업에 맞는 인공지능 반도체를 선택할 때 가장 큰 영향을 미친다. 이에 더불어 가장 적합한 반도체 선택을 위해서는 이외에도 메모리 용량, 유연성 등과 같은 다른 요소들 또한 균형 있게 고려해야 한다.

가. 연산 단위

연산 단위는 말 그대로 인공지능 반도체 상에서의 연산을 수행하는 기본 단위이며, 딥러닝 작업을 효

율적으로 수행하기 위해 설계되었다. 연산단위에 따라 대량의 데이터 및 복잡한 수학적 연산을 수행하므로 인공지능 반도체의 성능과 효율성을 결정짓는 가장 중요한 요소이다. 특히, 각 반도체마다 제공하는 연산 단위가 다르며, 특화된 연산 단위 또한 다를 수 있다.

나. 에너지 효율성

에너지 효율성은 인공지능 반도체가 소비하는 전력 대비 수행 가능한 연산의 양을 의미한다. 딥러닝 작업은 대량의 데이터 처리 및 복잡한 계산을 요구하기 때문에, 고성능을 유지하면서도 전력 소비를 최소화 하는 것이 필요하다. 주요한 전력소비 요인은 연산 자체와 더불어 외부 메모리에 접근하여 필요한 데이터를 가져오는 부분이다. DRAM 접근이 실제 연산에 비해 약 100배 이상의 전력이 소모되므로 DRAM에 접근하는 횟수를 줄여야한다. 이를 위해 저전력 설계 및 에너지 효율적인 연산 알고리즘들이 연구되고 있다. 에너지 효율적인 연산 알고리즘은 딥러닝 모델의 효율성 개선을 위해 필요하며, 연산량을 줄이면서 정확도를 유지 및 개선할 수 있어야 한다. 이를 위한 딥러닝 네트워크 아키텍처에 대한 연구들이 진행되고 있으며, 대표적으로 프루닝 (pruning) [5], 양자화 (quantization) [6], 그리고 지식 증류 (knowledge distillation) [7] 같은 기술이 있다.

다. 처리 속도

처리 속도는 모델의 학습과 추론 속도에 직접적인 영향을 미치기 때문에 뒤에서 언급될 TOPS 및 TFLOPS 등과 같은 인공지능 반도체의 성능 평가 지표에 영향을 준다. 빠른 처리 속도는 딥러닝 애플리케이션의 반응 속도를 향상시킬 수 있고, 복잡한 모델을 실시간으로 실행할 수 있게 하며, 대규모 데이터에 대한 학습과 추론을 가속화한다. 따라서, 처리 속도는 대규모 딥러닝 작업과 실시간 연산이 필요한 작업에서 특히 중요한 요소이다.

3. 국내외 인공지능 반도체 개발 동향

가. NVIDIA H100 [8,9]

NVIDIA의 H100은 해당 칩에 탑재된 Tensor Core로 인해 딥러닝 학습 및 추론, 고성능 컴퓨팅, 컴퓨터 비전 등에 매우 적합하다. 특히, 딥러닝 작업에 주로 사용되던 NVIDIA V100에 비해 연산 속도가 5배 이상 향상되었다. 또한 NVLink 스위치 시스템을 사용하면 최대 256개의 H100을 연결하여 작업을 가속화하고, 매개변수가 조 단위인 언어모델을 처리할 수 있다. 또한 여러 데이터 타입에 대한 연산을 지원하며 모든 데이터 타입에 대해 매우 높은 성능을 보인다. 이처럼 고성능의 실시간 추론 능력을 갖추었으나 이전 세대의 GPU보다 더 많은 전력 (W) 을 소모한다는 단점이 있다.

나. Meta MTIA [10]

Meta는 컨텐츠 추천을 위한 인공지능 모델을 적극적으로 활용하고 있으며, 이를 위해 MTIA라는 인공지능 반도체를 설계하였다. MTIA는 Meta Training and Inference Accelerator의 약자로 대규모 트레이닝 및 추론을 위한 인공지능 가속기이다. 해당 칩은 Pytorch를 지원하며, 해당 프레임워크에 완전히 최적화되도록 설계하였다. 또한, 128MB의 큰 On-Chip 메모리를 가지므로 자주 액세스하는 데이터 및 명령에 대해 더 낮은 대기시간이 요구된다. 또한, 전력 소모 대비 높은 성능을 갖는 칩으로 알려져있다. 또한 MTIA는 칩 하나로 학습과 추론을 모두 처리하게 했다는 점에서 주목을 받고 있다.

다. Tesla D1 [11]

Tesla에서는 자율 주행 자동차에 완전 자율 주행 기능 (Full Self Driving, FSD) 을 탑재하기 위해 인공지능 반도체인 D1을 개발하였다. 따라서 요즘 주목 받고 있는 GPT 등과 같은 거대 딥러닝 모델보다는 자율 주행과 연관된 딥러닝 모델을 대상으로 한다. 이를 위해서는 주로 비디오 데이터에 대한 학습 및 도로 위의 물체에 대한 자동 라벨링 등의 기능이 필요하다.

Tesla의 D1 프로세서는 354개의 DOJO 학습 노드를 통합한 On-Chip 네트워크 구조를 특징으로 하며, 각 DOJO 학습 노드 내부에는 BF16, FP8, INT8 등 다양한 데이터 타입을 처리할 수 있는 행렬 곱 전용 연산기와 벡터 연산기가 포함되어 있다. 이러한 구조를 통해 고도로 병렬화 된 연산이 가능해지며, 인공지능 학습 및 추론 과정에 필요한 복잡한 수학적 연산을 효율적으로 수행할 수 있도록 설계되었다.

라. Apple Neural Engine [12]

Apple Neural Engine (ANE) 은 Apple이 자체 개발한 신경망 엔진이며, 복잡한 기계 학습 모델을 빠르고 효율적으로 처리할 수 있도록 설계되었다. ANE는 2017년 iPhone X에 탑재된 A11 칩의 일부로 처음 포함되었다. A11은 최대 0.6 TFLOPS였지만, 2021년 5세대 16-core ANE는 26배가량 높아진 15.8 TFLOPS으로 성능이 향상되었음을 보여주었다. 이는 스마트폰에서도 높은 성능의 NPU를 사용할 수 있음을 나타낸다. M2 pro chip에는 CPU 및 GPU와 함께 딥러닝 모델의 교육 및 추론 프로세스를 가속화하기 위한 16-core Neural Engine이 탑재되었다. 따라서 M1 대비 연산 처리 속도가 40퍼센트 향상되어 초당 최대 15.8 TFLOPS의 연산을 처리할 수 있게 되었다. 이것은 다양한 기계 학습 기반의 작업을 실행하는 데 필요한 높은 처리 속도를 제공하는 것이며, 뛰어난 효율성과 반응 속도 또한 보장한다.

마. ETRI AB9 [13]

AB9은 FP16을 지원하고 이동통신, 자율주행, 지능형 로봇 및 드론 등에 활용하기 위해 개발되었다. 해당 칩은 인공지능 신경망의 추론 연산 가속화를 위한 System-On-Chip (SoC) 이다. AB9에는 연산을 가속화 하는 Systolic Tensor Core (STC) 를 활용하고 있다. STC에는 딥러닝의 핵심 연산인 MAC 연산뿐만 아니라 활성화 함수인 ReLU, 풀링 레이어의 한 종류인 maxpool 등과 같이 신경망의 레이어를 위해 필요한 대부분의 연산을 수행할 수 있도록 설계되어있다. 이처럼 추가 과정 필요 없이 신경망 처리 작업을 위한 프로그래밍이 가능하다는 특징을 가지고 있다. AB9의 개발 목적 상 전력 소비가 낮고 속도가 어느 정도 보장되어야 한다. 현재 FP16 기준으로 15W의 낮은 전력을 소모하면서 초당 40 TFLOPS라는 높은 연산 능력을 보여준다는 장점을 가지고 있다. 특히 자율주행 자동차에는 수백단위의 전력이 소모되면 발열 심하기 때문에 낮은 전력 소모가 필요하다. 따라서, 향후 10W 이내의 전력 소비량을 달성하여 자율주행 자동차 등에 안전하게 탑재되는 것을 목적으로 하고 있다.

바. SAPEON X220 / X330 [14]

국내 SK텔레콤에서 개발한 인공지능 반도체인 X220 또한 복잡한 딥러닝 모델을 학습시키고 추론할 수 있는 능력을 갖추고 있다. 이를 통해, 음성 인식, 이미지 분석, 자연어 처리등과 같은 여러 딥러닝 작업을 가속화할 수 있다. 이때, X220은 INT8 자료형만을 지원하는데, 이는 딥러닝 추론을 중점으로 두는 칩이기 때문이다. 또한, 전력 소모를 최소화하여 운영비용을 절감할 수 있도록 하는 것에 중점을 두었으며, 여러 딥러닝 프레임워크들 (Pytorch, Tensorflow) 과 호환되게 함으로써 딥러닝 기술의 대중화 및 상용화를 촉진시키고자 하고 있다.

이에 더불어 최근 SAPEON은 X330이 차세대 인공지능 반도체로 제안하였다. 해당 칩은 X220에서는 제공하지 않던 부동 소수점 연산 (FP16, FP8)을 지원하며, 기존의 정수 연산 단위에서는 INT8만 제공한다. 또한, 기존의 X220과 성능을 비교하였을 때, 4배 이상의 연산 성능 및 2배 이상의 전력 효율을 달성하였다. 또한, 2025년 공개를 목적으로 하고 있는 X430는 초거대 딥러닝 모델에 대한 학습을 목표로 하고 있다.

표 3 국내외 인공지능 반도체에 대한 비교 분석 (연산 단위, 연산 속도, 전력, On-Chip 메모리)

제품명	제조사	Data Types	TOPS	TFLOPS	W (Watt)	On-Chip Memory
H100	NVIDIA (국외)	FP64, TF32, FP32, FP16, FP8, INT8	3958 (INT8)	34 (FP64), 37 (FP32), 989 (TF32), 1979 (FP16), 3958 (FP8)	>700	116MB
MTIA	Meta (국외)	FP16, BF16, INT8	102.4 (INT8)	51.2 (FP16)	25	128MB
D1	Tesla (국외)	FP32, BF16, FP8, INT8	-	22.6 (FP32), 362 (BF16)	400	442.5MB
Neural Engine	Apple (국외)	FP16	-	15.8 (FP16)	-	-
AB9	ETRI (국내)	FP16	-	40 (FP16)	15~40	>36MB
X220	SAPEON (국내)	INT16, INT8, INT4	87 (INT8, Compact), 174 (INT8, Enterprise)	-	65 (Compact), 135(Enterprise)	-
X330		FP16, FP8, INT8	-	184 (FP16, Compact), 367 (FP8, Compact), 368 (FP16, Prime), 734 (FP8, Prime)	75~120 (Compact), 250 (Prime)	-
Warboy	Furiosa AI (국내)	INT8	64 (INT8)	-	40~60	32MB
ATOM	Rebellions (국내)	FP16, INT8, INT4, INT2	128 (INT8)	32 (FP16)	60~150	64MB

사. Furiosa AI Warboy [15]

Warboy는 대규모 인공지능 및 기계학습 워크로드를 가속화하기 위해 설계된 NPU이다. 즉, 복잡한 딥러닝 연산을 빠르게 처리할 수 있는 능력을 갖추었으며, 해당 칩을 기반으로 클라우드 컴퓨팅, 엣지 컴퓨팅 등과 같이 여러 환경에서의 인공지능 작업을 가속화 하는 데에 초점을 두고 있다.

해당 칩은 EfficientNetV2 (CNN 기반의 효율적인 대표적인 딥러닝 모델) 를 기준으로 NVIDIA의 T4와 비교하였을 때, 2.83배의 와트 당 처리량을 보였다. 이처럼 Warboy는 고도의 병렬 처리 아키텍처와 에너지 효율성을 고려하여 설계되었다. 이로 인해 대규모의 연산을 필요로 하는 딥러닝 작업을 수행할 때 발생하는 높은 전력 소모를 최소화하고, 양자화를 통해 처리속도 및 성능을 최적화하였다.

아. Rebellions ATOM [16]

ATOM은 최대 16개의 작업을 동시에 지원하도록 할 수 있으며, 효율적인 하드웨어 구현을 통해 좋은 성능을 제공한다. 해당 칩은 소규모 어플리케이션부터 대량의 데이터를 다루기 위한 작업에 이르기까지, 다양한 깊이 및 복잡성을 가진 네트워크를 효율적으로 가속화 할 수 있다. 또한, 통신 오버헤드 등과 같은 문제점을 극복하기 위해 에너지 효율성을 위한 저전력 설계 방법론을 적용한 것으로 알려져 있다.

특히, 해당 칩은 CNN, LSTM, BERT, 최신 트랜스포머 모델 (T5, GPTs 등)을 포함한 다양한 유형의 신경망을 효율적으로 가속화 할 수 있다.

4. 성능 비교 및 분석

본 절에서는 앞서 살펴본 국내외 인공지능 반도체들의 성능을 비교 분석한다. 이를 위해 TFLOPS (Tera Floating Point Operations Per Second)

와 TOPS (Tera Operations Per Second) 라는 성능 지표를 사용한다. 두 지표가 의미하는 바는 초당 테라 연산량이지만, 대상 연산의 종류가 다르다.

TFLOPS는 부동 소수점 연산에 대한 성능 지표에 해당하고, TOPS는 정수 연산까지 포함한 모든 종류의 연산에 대한 성능을 나타낼 수 있다. 인공지능 반도체에서는 부동 소수점 및 정수 연산이 모두 활용되는 경우가 많다. 그러나 각 칩마다 지원하는 연산 단위가 다르므로 본 논문에서는 두 지표를 혼용한다.

표 1에서 볼 수 있듯이 연산량과 전력은 트레이드오프가 존재하며, 이 두 요소를 모두 고려한 에너지 효율성을 평가하는 것이 중요하다고 생각된다. 또한, 향후 인공지능 반도체 개발에 있어 에너지 효율성을 향상시키기 위해 전력과 연산량의 균형을 맞추어가는 것이 중요한 부분이 될 것이다. 그러나, 현재로써는 대부분이 높은 연산 속도에 초점이 맞추어져 있다.

본 논문에서는 연산량, 전력 소모, On-Chip 메모리의 크기, 에너지 효율성에 초점을 맞추어 각 인공지능 반도체를 평가하고자 한다.

가. 연산 단위

연산 단위는 인공지능 반도체에서 매우 중요한 요소이다. 표 1에서 볼 수 있듯이 NVIDIA H100이 가장 많은 데이터 타입을 지원하여 가장 범용적인 연산이 가능하다. 이에 비해 다른 인공지능 반도체들은 거의 FP16, FP8, INT8을 지원하며, 부동소수점을 지원하지 않는 경우도 있다. 이처럼 특정 작업과 연산 단위에 특수화 된 반도체들이 존재한다.

나. 연산 속도

본 논문에서는 각 인공지능 반도체에 대한 연산 속도를 비교한다. 그러나 앞서 언급하였듯이 제공하는 연산 단위가 각각 다르므로 공정한 비교는 어려울 수 있다. 따라서 가장 많이 지원하는 연산 단위인 FP16과 INT8으로 나누어 연산 속도를 비교하였다. 해당 지표는 전력 소모나 범용성 및 특수성을 고려하지 않은 결과이다.

가장 빠른 연산 속도를 갖는 인공지능 칩은 H100이고, 그 다음으로 X330과 D1이 유사한 수준의 연산 속도를 보였다. X220, ATOM, MTIA가 뒤를 이었으며, AB9, Warboy는 비교적 낮은 연산 속도를 보였다.

다. 전력 소모

전력 소모는 각 인공지능 반도체가 연산을 할 때 소모되는 전력량이며, 단위는 W이다. 전력 소모는 인공지능 반도체에서 에너지 효율성을 위해 굉장히 중요한 요소이다. 해당 수치가 클수록 많은 에너지를 소비할 확률이 높음을 의미한다. 전력을 가장 많이 소모하는 칩은 H100이며, 다음으로는 D1, X330, ATOM, X220, Warboy 순으로 높은 전력 소모를 보였으며, 마지막으로는 AB9와 MTIA가 유사한 성능을 보였다. 즉, 가장 적은 전력을 소모할 수 있는 반도체는 AB9와 MTIA이지만, 위의 연산 속도에서 수 있듯이 빠른 연산속도를 제공하지는 않는다. AB9은 자율 주행을 위해 저전력 설계가 된 칩이며, D1은 자율주행에 사용되기에는 높은 소비 전력을 갖지만, 높은 연산 속도를 갖는다. 그러나 더 정확한 추론이 가능함을 주장하며 더욱 안전한 자율주행을 위해 설계되었다. 이처럼 개발 목적에 따라 전력에도 차이가 있음을 알 수 있다.

라. On-Chip 메모리

On-Chip 메모리는 말 그대로 칩 상에 존재하는 메모리이며, 외부의 DRAM에 접근하여 데이터를 가져온 후 인공지능 연산을 위해 저장해 놓는 메모리이다. 해당 메모리가 클수록 외부 메모리까지 접근할 필요 없이 많은 연산이 가능하다. DRAM에 접근하는 부분은 매우 큰 전력 소모를 야기하므로 On-Chip 메모리가 크면 일반적으로 지연 시간이 감소하고 연산 성능을 높일 수 있다. 그러나, On-Chip 메모리의 크기를 늘리는 것에는 추가 비용, 칩의 전력 소모 증가 등의 단점이 동반될 수 있으므로, 칩 설계 시 성능, 비용, 전력 소모량 사이의 균형을 찾는 것이 중요하다.

분석 대상 반도체 중, On-Chip 메모리가 가장 큰 순서대로 분석 대상 인공지능 반도체를 나열하면, D1, MTIA, H100, ATOM, AB9, Warboy 순이다. 이외의 칩은 제공된 정보가 없으므로 제외되었다.

마. 에너지 효율성

에너지 효율성은 인공지능 반도체가 소비하는 전력 대비 수행 가능한 연산의 양을 의미한다. 연산 단위가 달라 모든 AI 칩에 대한 공정한 비교가 어려우므로 연산 속도와 마찬가지로 가장 많이 사용되는 FP16과 INT8에 대해 계산하였다. 에너지 효율성의 정의에 따라 TOPS 또는 TFLOPS를 전력 소모량(W) 로 나누었다 (TOPS or TFLOPS per W). 따라서 결과 값이 클수록 에너지 효율성이 높음을 의미한다.

FP16에 대한 에너지 효율성이 높은 순서로 나열하면 다음과 같다: H100 (2.82), AB9 (2.66~1.00) 및 X330 (2.45~1.53). MTIA (2.05), D1 (0.91), ATOM (0.53~0.21). FP 16에 대한 결과로, H100이 높은 전력에도 불구하고 높은 에너지 효율성을 보였다.

다음으로 INT8에 대한 에너지 효율성이 높은 순서대로 나열하면 다음과 같다: H100 (5.65), MTIA (4.10), ATOM (2.13~0.85) 및 X220 (1.33, 1.28), Warboy (1.06). 따라서 H100이 에너지 효율성에서 가장 좋은 성능을 보인다.

5. 기존의 하드웨어 및 소프트웨어 최적화 방법론에 대한 논의

위에서 분석한 인공지능 반도체의 연구 현황 및 각 칩에 대한 연산 단위, 속도, 에너지 효율성 등을 바탕으로 향후 연구 방향을 제시한다. 현재 연구되고 있는 고성능 인공지능 반도체에 대한 연구도 계속하여 진행되어야 하지만 각 용도에 맞는 효율적 설계 또한 중요한 부분이 될 것이다. 인공지능 반도체의 효율성과 고속화를 위한 요소들은 반도체의 하드웨어 아키텍처와 알고리즘의 최적화 수준과 관련이 있다. 하드웨어와 소프트웨어 간의 최적화를 통해 전반적인 시스템 성능을 향상시키는 데 중점을 두는 향후 연구들이 수행되어야 할 것이다.

가. 인공지능 반도체의 아키텍처

인공지능 반도체의 고속화를 위해서는 특정 계산에 대한 병렬 처리에 최적화 된 아키텍처를 갖춤으로써 많은 연산을 빠르고 효율적으로 수행할 수 있어야 한다. 또한, NVIDIA의 Tensor Core와 같이 특정 데이터 타입과 연산 대상이 되는 텐서의 특징 (Density, Sparse 행렬 등) 에 따라 행렬 곱셈 연산을 가속화 하도록 설계하면, 딥러닝 연산의 대부분을 차지하는 연산이 가속화 되므로 모델의 성능을 크게 개선할 수 있다. 마지막으로, 거의 모든 칩에서 통신 과정에서의 오버헤드가 발생하는 경향이 있다. 이를 위해 향후에는 컴퓨팅 성능 향상, 소비 전력 최소화, 메모리 대역폭 및 통신 오버헤드 관리 간의 균형을 유지하는 데에 초점을 두어야 할 것으로 보인다.

나. 소프트웨어 및 알고리즘 최적화

하드웨어가 아닌 소프트웨어 및 알고리즘을 최적화하여 인공지능 반도체의 효율성을 향상시킬 수 있다. 소프트웨어 최적화는 하드웨어의 성능을 최대화하기 위한 소프트웨어 및 알고리즘의 최적화를 의미하며, 앞서 언급한 pruning, quantization, zero value skipping 등과 같은 기술을 통해 계산 부하를 줄이고 메모리 사용량과 에너지 소비를 최소화할 수 있다.

1) Zero value skipping

해당 방식은 convolutional neural network (CNN) 에서 효율적인 최적화 알고리즘이다. 이는 연산 대상인 텐서에서 0인 값을 제거함으로써 연산 및 전력 소모를 감소시키는 방법이다. 연산 과정에서 0이 곱해지면 누적 연산 시 영향이 가지 않기 때문에, 존재하지 않는 값이나 마찬가지이다. 따라서 값이 0인 텐서 요소들을 제거하면, 연산에 필요한 사이클을 감소시킬 수 있게 된다, Convolution layer의 연산 중 37~50%는 0 값을 곱하는 연산이고, 이와 더불어 활성화 함수인 ReLU는 음수에 대한 값을 0으로 출력하기 때문에 zero skipping 알고리즘은 딥러닝 가속기 상에서의 연산에 있어 중요한 최적화

알고리즘이다. 이와 관련하여 실제로 zero skipping을 위한 'Cnvlutin'이라는 가속기 [17]가 존재한다. 그러나, 0 값을 선택하여 제거하기 위해서는 index 개념의 offset이 사용되어야 한다. 이 과정에서 offset을 저장하기 위해 필요한 메모리가 증가하게 된다. 그러나 전체적으로 볼 때, convolution layer에서의 연산 사이클이 굉장히 줄어들기 때문에 에너지 측면에서의 절전 효과가 존재한다.

2) Pruning

해당 방식은 딥러닝 연산에서 가중치 가지치기라고 불리는 최적화 알고리즘이다. 네트워크에 영향을 적게 주는 가중치 (즉, 임계치 이하의 가중치)를 배제하기 위해 0으로 간주하고 연산하는 방식이다. 이를 통해 연산 대상이 되는 노드 자체를 줄여버림으로써 연산량 감소의 효과를 얻을 수 있다.

3) 명령어 세트 아키텍처 (Instruction Set Architecture, ISA)

인공지능 작업을 위한 ISA에 대한 연구도 진행되었다 [18]. 딥러닝 연산을 제공하는데 과도한 하드웨어 리소스를 사용하도록 하는 경우가 있으므로 명령어 수준에서의 유연성을 확보할 필요가 있다. 이는 각 모델에 대한 명령어가 아니라 여러 모델에 대해 범용적으로 사용될 수 있다면 좋을 것으로 생각된다.

즉, 인공지능 반도체의 효율성 및 고속화를 위해서는 하드웨어뿐만 아니라 효율적인 소프트웨어 요소들이 뒷받침되어야 한다.

다. 에너지 효율성 및 고속화를 위한 향후 연구 방향 제시

본 절에서는 앞서 논의한 인공지능 반도체의 특징과 효율성 향상 및 고속화 방법을 기반으로 향후 연구 방향을 제시한다. 가장 먼저, 하드웨어와 소프트웨어 간의 최적화이다. 딥러닝 알고리즘과 인공지능 반도체 설계의 상호 작용을 극대화함으로써 주로 사용되는 데이터 타입이나 특성 그리고 행렬 연산을 고려하여 전반적인 성능을 향상시키기 위한 연구들이 활발히 진행되고 있으며, 앞으로도 지속되어야한다.

다음으로는 이러한 성능 향상과 더불어 이제는 에너지를 효율적으로 소모하기 위한 여러 연구들이 필요하다. 현재 대부분의 인공지능 반도체들이 높은 성능을 대상으로 하고 있으나, 해당 반도체들이 사용되는 용도가 모두 다르기 때문에 활용 목적에 맞는 에너지 효율성을 갖춰나가야 할 것이다.

다음은 하드웨어 유연성 확보이다. 앞서 살펴보았듯이, 현재는 특정 작업에 최적화 된 칩들이 존재한다. 그러나 이는 특정 작업에 최적화 된 명령어를 사용함으로써 범용성이 떨어지기 때문에, 다양한 딥러닝 작업에 사용할 수 있는 유연성을 갖춘 반도체의 개발과 소프트웨어와의 호환성이 중요해질 것으로 생각된다. 이를 위해, 레이어 별 연산 제공 및 사용 시간에 따른 최적 연산 재구성 등이 가능한 아키텍처에 대한 연구가 필요할 것으로 보인다.

마지막으로는 인공지능 작업의 특성상 개인 정보와 같은 민감 데이터를 처리하는 데에 널리 활용되고 있으므로, 보안 및 프라이버시 보호 기능을 내장한 설계 또한 중요해질 것으로 생각된다.

Ⅲ. 결 론

본 논문에서는 인공지능 반도체의 정의와 국내외 인공지능 반도체의 연구 동향에 대해 살펴보고, 향후 연구 방향을 제시한다.

국내외 인공지능 반도체 연구 동향에 대해 분석한 결과 단순히 높은 연산 속도만이 중요한 것이 아니라 제공하는 데이터 타입과 전력 소모량 그리고 On-Chip 메모리의 크기 등이 전체 연산 효율에 영향을 미침을 확인하였다. 현재로서는 NVDIA의 H100이 여러 요소에서 가장 선두에 있으며, D1이나 AB9와 같은 칩은 특정 목적에 맞게 설계하기 위한 연구를 진행하였다.

향후에는 연산 성능 향상과 더불어 목적에 맞는 칩 설계, 에너지 효율성 향상, 하드웨어 및 소프트웨어 간의 최적화 등에 초점을 맞춘 연구가 진행되어야 할 것으로 생각된다.

REFERENCES

[1] Owens, John D., et al. "GPU computing," *Proceedings of the IEEE 96.5,* pp. 879-899, 2008.

[2] Choquette, Jack, et al. "NVIDIA A100 tensor core GPU: Performance and innovation," *IEEE Micro,* vol. 41, no. 2, pp. 29-35, 2021.

[3] Wang, Yu Emma, Gu-Yeon Wei, and David Brooks. "Benchmarking TPU, GPU, and CPU platforms for deep learning," arXiv preprint arXiv:1907.10701, 2019.

[4] Choi, Yujeong, and Minsoo Rhu. "Prema: A predictive multi-task scheduling algorithm for preemptible neural processing units," *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE,* 2020.

[5] Hoefler, Torsten, et al. "Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks," *The Journal of Machine Learning Research 22.1,* pp. 10882-11005, 2021.

[6] Wu, Hao, et al. "Integer quantization for deep learning inference: Principles and empirical evaluation," arXiv preprint arXiv:2004.09602, 2020.

[7] Gou, Jianping, et al, "Knowledge distillation: A survey," *International Journal of Computer Vision 129:,* pp. 1789-1819, 2021.

[8] NVIDIA H100 Tensor Core GPU Architecture: EXCEPTIONAL PERFORMANCE, SCALABILITY, AND SECURITY FOR THE DATA CENTER (2022), https://www.advancedclustering.com/wp-content/uploads/2022/03/gtc22-whitepaper-hopper.pdf, (accessed 18, 03, 2024)

[9] Anne C. Elsterm et al. “Nvidia Hopper GPU and Grace GPU Highlights,” *Computing in Science & Engineering,* vol. 24, no. 2, pp. 95-100, 2022.

[10] MTIA v1: Meta’s first-generation AI inference accelerator (2023), https://ai.meta.com/blog/meta-training-inference-accelerator-AI-MTIA, (accessed 18, 03, 2024)

[11] E. Talpes et al., “The microarchitecture of dojo, tesla’s exa-scale computer," *IEEE Micro*, vol. 43, no. 3, pp. 31-39, 2023.

[12] Deploying Transformers on the Apple Neural Engine (2022), https://machinelearning.apple.com/research/neural-engine-transformers, (accessed 18, 03, 2024)

[13] Yong Cheol Peter Cho, et al.. “AB9: A neural processor for inference acceleration,” *ETRI ETRI Journal,* vol. 42, no. 4, pp. 491-504, Aug. 2020.

[14] Sapeon (2024), https://www.sapeon.com/, (accessed 18, 03, 2024)

[15] FuriosaAI (2024), https://furiosa.ai/warboy/specs, (accessed 18, 03, 2024)

[16] ATOM: 5nm Versatile Inference SoC, Versatile yet Energy Efficient AI System-on-Chip (2023), https://rebellions.ai/rebellions-product/atom-2/, (accessed 18, 03, 2024)

[17] Albericio, Jorge, et al. "Cnvlutin: Ineffectual-neuron-free deep neural network computing," *ACM SIGARCH Computer Architecture News,* vol. 44, no. 3 pp. 1-13, 2016.

[18] Liu, Shaoli, et al. "Cambricon: An instruction set architecture for neural networks," *ACM SIGARCH Computer Architecture News,* vol. 44, no. 3, pp. 393-405, 2016.

저 자 소 개

김 현 지(학생회원)

2020년 2월: 한성대학교 IT 응용시스템 공학과 학사 졸업.
2022년 2월: 한성대학교 IT융합공학과 석사 졸업.
2023년 3월: 한성대학교 정보컴퓨터공학과 박사과정.
<주관심분야 : 정보보안, 인공지능, 양자 컴퓨팅>

윤 세 영(학생회원)

2024년 2월: 한성대학교 IT 융합공학부 학사 졸업.
2024년 3월: 한성대학교 융합보안학과 석사 과정.
<주관심분야 : 인공지능, 디지털 포렌식>

서 화 정(정회원)

2010년 2월: 부산대학교 컴퓨터공학과 학사 졸업.
2012년 2월: 부산대학교 컴퓨터공학과 석사 졸업.
2016년 2월: 부산대학교 컴퓨터공학과 박사 졸업.
2017년 3월: 싱가포르 과학기술청
2023년 2월: 한성대학교 IT융합공학부 조교수
2023년 3월: 한성대학교 융합보안학과 부교수
<주관심분야 : 암호구현, 정보보안>