ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 52, No. 12, pp. 1009-1015, 2025. 12
https://doi.org/10.5626/JOK.2025.52.12.1009

CNN 모델에 대한 임베디드 가속기 플랫폼의 성능 및 비용분석

(Performance and Cost Analysis of an Embedded Accelerator Platform for CNN Workloads)

권 도 원 †　　구 건 재 ††
(Dowon Kwon)　　(Gunjae Koo)

요 약 최근 인공지능 어플리케이션은 다양한 시스템에서 구동되고 있으며 이를 효율적으로 처리하기 위하여 하드웨어/소프트웨어적인 최적화 방법이 적용되고 있다. 그러나 제한된 자원을 가진 임베디드 환경에서는 범용 컴퓨팅 환경에 적용되는 최적화 방법을 바로 적용하기 어렵기 때문에 임베디드 환경에 특화된 성능 및 비용 분석이 필요하다. 본 연구는 임베디드 환경에서 범용 가속기(GPU)와 AI 특화 가속기(DLA)를 비교하고 상황에 따라 적합한 가속기를 선택할 수 있는 기준을 제시한다. 실험결과 CNN 모델에서 DLA는 GPU에 비해 최대 4.5배 정도 전력 효율성을 보였다. 반면 처리량, 지연시간, 에너지 효율 측면에서는 GPU보다 낮은 성능을 기록하였다. 종합적으로 임베디드 환경에서 AI 특화 가속기는 전력 효율 면에서 우수한 성능을 나타내고 있으며 범용적인 워크로드들을 동시에 처리할 때 CPU/GPU와 같은 범용 프로세서와 병렬로 분산하여 활용할 때 성능 및 비용의 최적화가 가능하다.

키워드: 임베디드 플랫폼, GPU, 가속기, CNN 모델, 성능 분석

Abstract Modern artificial intelligence (AI) applications operate on a wide range of systems. Researchers have presented hardware and software optimizations to process AI applications efficiently. However, in embedded systems with limited resources, it is challenging to directly apply the optimizations designed for general-purpose computing systems. This study compares general-purpose accelerators (GPUs) and AI-specialized accelerators (DLAs) in embedded platforms to provide criteria for selecting the appropriate processors for convolutional neural network (CNN) models. Our experiment results show that DLAs demonstrate up to 4.5 times higher power efficiency compared to GPUs. On the other hand, DLAs exhibit lower performance in terms of throughput, latency, and energy efficiency. Overall, this study presents that CNN models can be optimized for different target embedded processors based on the constraints of embedded system environments.

Keywords: embedded AI platform, GPU, accelerator, CNN models, performance analysis

· 이 성과는 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1C1C1012172)

† 비 회 원 : 고려대학교 컴퓨터학과 학생
dnjs1510@korea.ac.kr
†† 정 회 원 : 고려대학교 컴퓨터학과 교수(Korea Univ.)
gunjaekoo@korea.ac.kr
(Corresponding author임)

논문접수 : 2024년 11월 14일
(Received 14 November 2024)
논문수정 : 2025년 8월 19일
(Revised 19 August 2025)
심사완료 : 2025년 9월 7일
(Accepted 7 September 2025)

Copyright©2025 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 제52권 제12호(2025. 12)

1. 서 론

인공지능의 발전으로 인하여 인공지능 어플리케이션이 요구하는 대용량의 데이터를 병렬적으로 빠르게 처리하기 위한 처리 장치들의 개발 경쟁이 활발하다. 또한 인공지능을 실행하는 플랫폼도 고성능 서버 시스템으로부터 임베디드 시스템까지 다양해지고 있다. 인공지능 어플리케이션이 수행되는 시스템의 특성에 따라 우선순위가 되는 성능 지표가 상이하다. 예를들어, 대량의 데이터를 처리하는 데이터 센터는 처리량과 속도가 중요하지만 임베디드 시스템은 전력, 크기, 발열과 같은 물리적 제한 조건을 가지고 있어 이를 고려한 처리 장치 선택이 필요하다.

본 연구는 여러 플랫폼 중에서 임베디드 시스템의 AI 가속기에 집중했다. 일반적으로 병렬적인 인공지능 데이터를 처리하는 처리장치에는 GPU(graphics processing unit)가 대표적이다. 모든 플랫폼에 고성능 GPU를 탑재하는 것이 성능 향상과 호환성의 균형에 최선이지만 GPU의 성능은 SM(streaming multiprocessor)의 수에 비례한다. SM은 내부의 CUDA 코어나 캐시 등으로 구성되어 그 수가 많아질수록 GPU의 병렬처리 능력이 향상되지만 크기와 발열, 전력 소모가 증가하는 문제가 있다. 따라서 고성능 GPU를 임베디드 시스템에 적용하는 것은 물리적인 한계가 있다. 이를 해결하기 위해 기업들이 임베디드 환경에 적합하고 특정 어플리케이션에 최적화된 NPU, TPU, DLA와 같은 AI 전용 가속기를 개발하고 있다.

임베디스 시스템에 탑재된 가속기를 평가하는 기준은 다양하다. 일반적으로 전력 소모량, 에너지 효율성을 평가하고 비교 분석한다. 본 연구는 임베디드 플랫폼에 탑재된 AI 가속기를 활용해 CNN 모델에 대한 워크로드 분석을 전력 소모, 에너지 효율성뿐만 아니라 운용 안정성 측면에서도 검토하고 Nvidia Jetson AGX Xavier에 내장된 GPU와 DLA를 동일한 조건에서 비교하고 두 가속기 간의 성능 차이를 구조적 특성에 중점을 두고 설명할 예정이다.

본 연구에서 진행한 실험은 Jetson AGX Xavier의 GPU와 AI 전용 가속기인 DLA로 한정하며, 동일한 실험환경 하, CNN 모델 추론 과정에서 발생하는 두 가속기의 성능 차이를 분석했다. 실험을 진행한 Jetson의 GPU는 기존 GPU와는 달리 임베디드 시스템에 적합한 저전력 기반의 GPU며, DLA는 Jetson 플랫폼에서 인공지능 연산에 필요한 행렬 연산에 특화된 특징을 가지고 있다[1]. 따라서 GPU와 DLA는 다른 아키텍처와 치적화 방식을 기반으로 하며, 이에 따른 성능 차이가 발생하는 이유를 각 가속기의 특성을 중심으로 분석하고 추론 과정에서 발생하는 전력 소모량, 에너지 효율성과 운용의 안정성을 평가하고 성능을 비교하였다.

본 논문의 구성은 다음과 같다. 먼저 임베디드 AI 가속기와 관련된 연구 현황을 살펴보고 Nvidia Jetson AGX Xavier의 특성과 내장된 AI 가속기를 설명한다. 이후 두 가속기의 구조적인 특성과 성능 차이 분석을 다루고, 실제 응용 환경에서의 차이와 향후 연구 방향을 논의한다.

2. 관련연구

인공지능 워크로드를 분석해 가속기 성능을 평가하는 연구는 크게 두 부분으로 나눌 수 있다. 첫째, GPU 구조에 기반한 AI 연산에 최적화된 가속기 개발이 있으며, 이는 범용 가속기인 GPU의 구조를 AI 연산방식에 적합하게 개선하는 것으로 시작되었다. 이후 NPU나 TPU와 같은 AI 특화 가속기가 개발되면서 성능과 효율성이 더욱 향상되었다. 둘째, 인공지능 모델 경량화, 양자화, 프루닝과 같은 기법을 통해 워크로드를 최적화하는 연구도 활발하다. 이러한 방법은 모델 크기를 축소하거나 처리 속도를 향상시키고 최적화를 통해 전체적인 효율성을 높이는 데 기여한다[2-6].

본 연구는 관련 연구를 바탕으로 범용 가속기(GPU)와 AI 전용 가속기(DLA)에서 AI 워크로드를 실행하여 성능을 비교하고 다양한 성능 평가 기준에 따라 최적의 가속기를 선정하고자 한다.

3. 배 경

이 장에서는 Nvidia Jetson AGX Xavier의 전체적인 구조를 설명하고 범용 가속기(GPU)와 AI 전용 가속기(DLA) 구조에 대해 설명한다.

3.1 Nvidia Jetson AGX Xavier

임베디드 인공지느 플랫폼은 인공지능 어플리케이션을 제한된 환경인 임베디드 시스템에 적용하기 위해 최적화된 방법을 사용하고 있다. 본 연구에 활용한 Jetson AGX Xavier는 대표적인 임베디드 인공지능 플랫폼으로 저전력 GPU 뿐만 아니라 딥러닝 연산에 최적화된 DLA(deep learning accelerator)가 탑재되어 있어 임베디드 환경에서도 효율적으로 딥러닝 워크로드를 처리할 수 있다[5].

Jetson AGX Xavier는 8개의 ARM v8.2 64-bit CPU 코어, 512 CUDA 코어와 64개의 Tensor 코어를 갖춘 Volta GPU, 2개의 DLA, PVA, 그리고 32GB LPDDR4x 메모리로 구성되어 있다[1,7](그림 1).

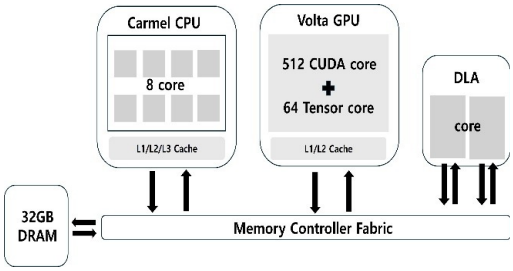

그림 1 Nvidia Jetson AGX Xavier 시스템 구성요소
Fig. 1 Nvidia Jetson AGX Xavier System Architecture

3.2 Volta GPU

Jetson AGX Xavier에 탑재된 Volta GPU는 CUDA 코어, Tensor 코어와 최적화된 메모리 구조로 인해 딥러닝 연산에 최적화되어 있으며 기존 GPU에 비해 전력 효율이 약 1.5배에서 2배 정도 향상되었다[2]. Volta GPU는 8개의 SM으로 구성되어 있으며 각 SM은 64개의 CUDA 코어와 tensor 연산을 위한 8개의 Tensor 코어를 갖추고 있다. 또한 각 SM은 128 KB의 L1 캐시를 가지고 있으며, 512KB의 L2 캐시가 각 SM간 데이터 공유를 지원한다[1,8](그림 2).

3.3 DLA

특정 AI 연산을 가속화하기 위해 플랫폼 마다 다양한 가속기를 탑재한다. Nvidia Jetson AGX Xavier는 딥러닝 연산 중 CNN(convolution neural network) 모델에 특화된 DLA가 탑재되었다. DLA는 CNN 모델의 추론 과정을 최적화하며 2개의 DLA 코어는 각각 독립적으로 동작하여 서로 다른 모델을 병렬 처리할 수 있다[8,9].

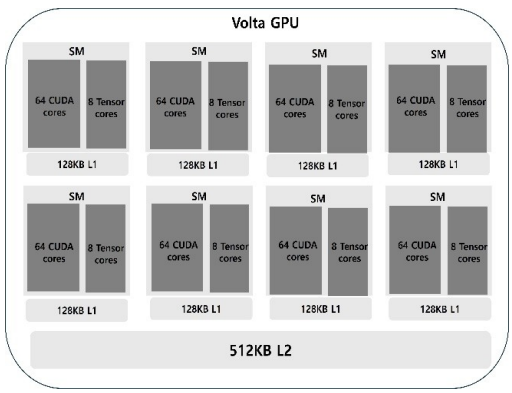

그림 2 Volta GPU 구조
Fig. 2 Volta GPU Architecture

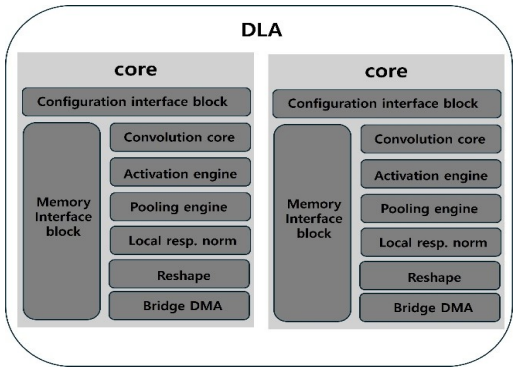

그림 3 DLA 구조
Fig. 3 DLA Architecture

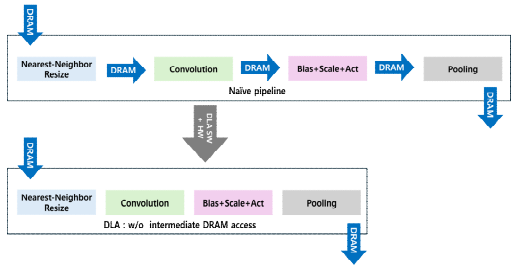

그림 4 GPU와 DLA의 메모리 접근 방식
Fig. 4 Memory Access Models of GPU and DLA

DLA는 convolution 코어, single data point processor (SDP), planar data processor(PDP), cross-channel data processor, data reshape engines, bridge DMA로 구성되어 있다. CNN 모델은 convolution(CONV), activation(ACT), pooling(POOL) 그리고 fully-connected (FC) 레이어로 구성된다. 따라서 DLA 구조 자체가 보편적인 CNN 모델에 최적화 되어있다(그림 3). 또한 각 DLA 코어는 512KB의 on-chip memory를 사용해 연산에 필요한 데이터를 저장하고 빠른 접근을 통해 GPU 대비 오버헤드를 최소화한다(그림 4).

3.4 GPU와 DLA 비교

GPU는 범용 컴퓨팅 가속기(general purpose compute accelerator)로서 병렬화가 가능한 다양한 워크로드를 처리할 수 있다. 구조적으로 수많은 SM이 동시에 많은 데이터를 처리하며 SIMT(single instruction, multiple threads) 패러다임을 통해 다수의 쓰레드가 병렬 연산을 가속화한다. 이런 특성으로 인해 데이터 전처리, 이미지 변환과 같은 연산량이 큰 작업에 적합하며 광범위한 연산을 요구하는 고성능 작업에서 우수한 성능을 나타낸다. 반면 DLA는 CNN 기반 추론 작업에 최적화된

표 1 GPU, DLA 특성 비교
Table 1 Comparison of GPU and DLA Characteristics

Characteristics	GPU	DLA
Type	General Purpose Computing Accelerator	Deep Learning Inference Accelerator (CNN)
Operation Method	SIMT	Specialized Operation
Optimized Task	Diverse parallel workloads and computationally intensive high-performance tasks	CNN-based deep learning model inference tasks
Flexibility	High flexibility, general purpose	Low flexibility (specialized for inference tasks)

가속기이며, 구조적으로는 convolution 및 activation 연산에 최적화되어 있다.

4 CNN 모델 최적화

Jetson에 탑재된 Volta GPU에 CNN 모델을 최적화하는 주요 방법은 TensorRT를 활용하는 것이다. TensorRT는 모델을 GPU에 맞춰 압축하고 불필요한 레이어를 제거하여 처리 속도를 높이고 메모리 효율성을 개선한다.

DLA는 구조적으로 범용 CNN 모델에 최적화되어 있다. 특히 convolution 연산을 빠르고 효율적으로 처리하는 데 적합하다. 다만 DLA에서 모델을 실행하기 위해서는 먼저 TensorRT를 사용해 모델을 최적화해야 한다. 그러나 DLA는 모든 연산이나 레이어를 처리하는 데 제약이 있을 수 있어 이러한 경우 Jetson 플랫폼에서는 TensorRT 최적화 과정에서 해당 연산을 GPU로 자동 할당하여 성능 저하 없이 최적화를 유지한다.

4.1 DLA 연산 파이프라인

DLA는 CNN 모델의 연산에 효율적으로 동작할 수 있도록 설계되었고 이 과정에서 연산 순서를 최적화하여 성능을 극대화한다. 연산이 이루어지는 동안 시스템 메모리에 반복적으로 접근하지 않고 데이터를 한 번만 전송해 메모리 접근을 최소화한다. 이런 방식으로 데이터 전송의 비효율성을 줄이고 연산 속도를 높여 결론적으로 전력 소비를 줄이고 성능을 높이는 데 유리하다.

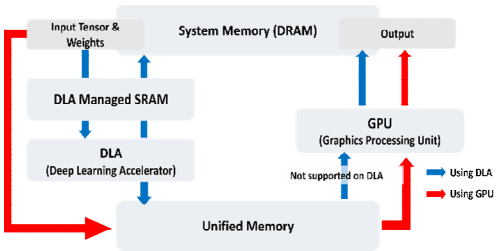

그림 5 GPU/DLA 최적화 과정
Fig. 5 GPU/DLA Execution Flow

4.2 최적화 결과

TensorRT를 사용해 CNN 모델을 GPU와 DLA에 적합한 형태로 모델을 최적화했다. 이 과정에서 가속기 특성에 맞게 각 레이어는 압축되고 최적화된다. 그 결과 DLA를 사용하는 경우 GPU보다 더 많은 레이어를 압축하고 최적화했다. GPU의 최적화율은 90%인 반면, DLA는 97%의 최적화율을 보였다. 최적화 작업에 대해 CNN 모델에서는 DLA가 더 효율적이지만, CNN 외 모델에서는 GPU가 상대적으로 더 높은 효율성을 보인다[10].

4.3 DLA 최적화 모델 특성

딥러닝 모델의 연산 과정에서 각 모델 구조적 특성에

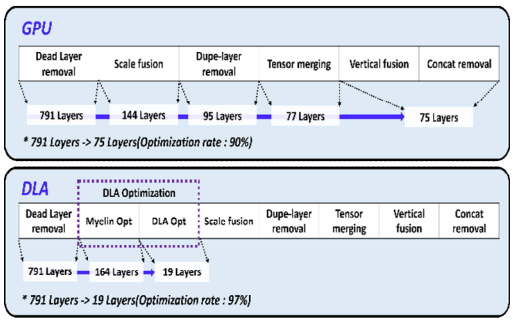

그림 6 딥러닝 모델 최적화 과정 및 결과
Fig. 6 Deep Learning Model Optimization

표 2 GPU/DLA Layer 할당 결과
Table 2 GPU/DLA Layer Assignment Breakdown

Model	GPU(%)	DLA(%)
AlenNet	20.0	80.2
Inceptionv1	15.0	85.0
Inceptionv3	30.0	70.0
GoogleNet	10.0	90.0
VGG16	40.0	60.0
Super Resolution	20.0	80.0
YOLOv3	30.0	70.0
YOLOv3-Tiny	20.0	80.0
ResNet18	20.0	80.0
ResNet34	20.0	80.0
ResNet50	20.0	80.0
Mobilenetv2	20.0	80.0

따라 일정 비율로 나뉘어 연산 및 레이어가 가속기에 할당된다. 또한 컨볼루션 레이어 비중이 높은 모델의 경우 DLA에 더 많은 연산이 할당되는 경향이 있다.

5. 임베디드 가속기 플랫폼 성능 및 비용 분석

본 연구에서는 다양한 CNN 모델을 벤치마크해 Nvidia Jetson AGX Xavier의 GPU와 DLA의 성능 및 비용을 분석하였다. 이 비교는 다양한 성능 지표를 활용했지만 임베디드 플랫폼에 우선시되는 전력 효율성과 운용 안정성을 중점적으로 두 가속기의 성능을 평가하는 데 초점을 맞추었다. 실험에 사용한 CNN 모델은 이미 학습된 상태에서 추론 작업만 실행했으며, 각 가속기가 지원하는 FP16과 INT8 두 가지 데이터 정밀도를 기준으로 실험을 수행했다. 이를 통해 정밀도에 따른 가속기 성능 차이 또한 평가했다[11]. 표 3은 실험에 사용된 플랫폼의 설정값을 보여준다. 전력 측정은 Jetson 플랫폼에서 제공하는 전력 측정 센서값을 사용하였다.

5.1 전력 소비량

실험 결과 DLA는 모든 CNN 모델에서 GPU보다 전력 소모가 낮은 것으로 나타났다. 또한 FP16보다 INT8 양자화를 적용한 경우 연산 량의 감소로 DLA의 전력 소모가 감소했다(그림 7). GPU 대비 저전력으로 설계되었으며 구조적으로 CNN 모델에 최적화되어 있어 낮은 전력 소모를 보인다. 반면 대량의 데이터를 처리하고 병렬 작업에 특화된 GPU의 경우 연산 과정에서 많은 전력을 소모하는 것으로 확인했다.

표 3 실험 환경
Table 3 Experiment Configuration

Platform	Nvidia Jetson AGX Xavier(32GB)
OS	Ubuntu 20.04
CUDA ver	11.4
Power Mode	MANX
GPU/DLA Freq	1377 MHz / 1395.2 MHz
Iterations	1,000

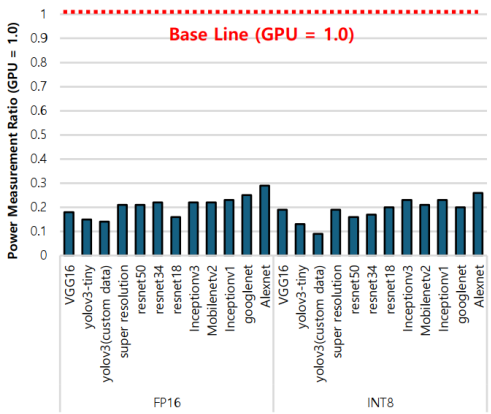

그림 7 DLA 전력 소모 (GPU 대비)
Fig. 7 Power Consumption by DLA (Normalized to GPU)

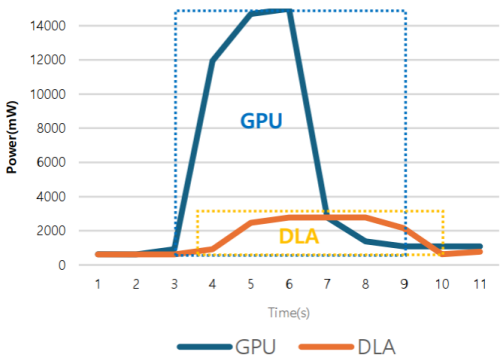

그림 8 전력 소비량 변화
Fig. 8 Power Consumption by Time

또한 추론 과정에서 두 가속기의 전력 소비량 변화를 관측할 때 GPU 단독으로 추론 시 순간 전력이 10,000mW 이상으로 급증하는 현상이 관측됐다. 반면 DLA는 전력 소비가 보다 일정하게 유지되어 시스템 불안정성을 줄일 수 있음을 확인했다(그림 8).

5.2 지연시간

지연 시간은 전력 소모와는 반대되는 경향을 보였다. GPU의 여러 SM을 통해 대규모 데이터를 병렬로 처리할 수 있어 DLA보다 빠른 처리 속도를 나타냈다. 특히

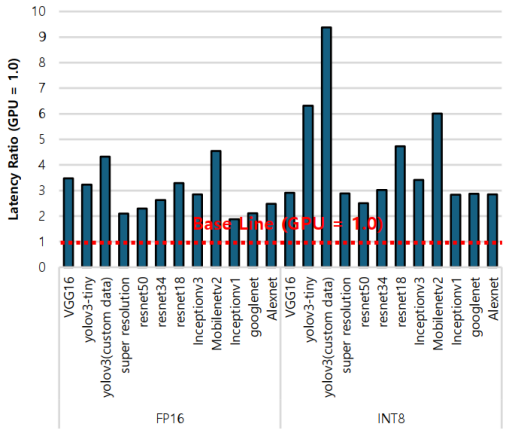

그림 9 DLA 지연시간 (GPU대비)
Fig. 9 Latency by DLA (Normalized to GPU)

MAC 연산이 많은 모델(예: Yolo v3)에서는 GPU와 DLA 간의 지연시간 차이가 더욱 두드러지게 나타났다. 이는 DLA가 MAC 연산이 많은 모델의 파라미터와 특성에 따라 연산 속도가 영향을 받는다는 것을 시사한다(그림 9).

5.3 처리량

처리량 측면에서 GPU는 DLA보다 월등히 높은 성능을 보였다. FP16보다 INT8 양자화 시 GPU는 처리량이 증가했지만 DLA의 경우 일부 모델에서 양자화 했을 때 처리량이 오히려 감소하는 경향이 관찰되었다(그림 10).

5.4 에너지 소비

에너지는 전력 소모가 지속적으로 일어나는 시간의 합을 나타낸다. DLA는 GPU 보다 낮은 에너지 소비를 기록했는데 이는 지연 시간이 다소 증가했지만 상대적으로 적은 전력 소모 덕분이다. FP16보다 INT8로 양자화 했을 때도 에너지 소비가 더 감소했다(그림 11).

5.5 전력 및 에너지 효율

전력 효율(power efficiency = throughput/power)은 DLA가 GPU보다 월등히 높은 것으로 나타났다. 모든 모델에서 DLA는 더 높은 전력 효율을 보였으며, 두 가속기 모두 INT8 양자화 시 더 나은 전력 효율을 기록했다(그림 12).

그러나 에너지 효율 (energy efficiency = throughput/energy)은 모델에 따라 다른 결과를 보였다. 단순한 convolution 연산을 반복하는 구조의 모델인 GoogleNet, AlexNet에서 DLA는 더 높은 에너지 효율을 기록한 반

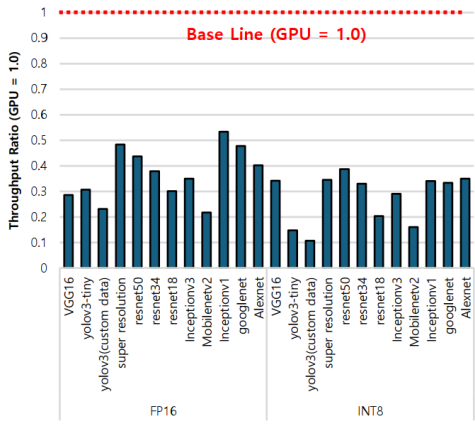

그림 10 DLA 처리량 (GPU대비)

Fig. 10 Throughput by DLA (Normalized to GPU)

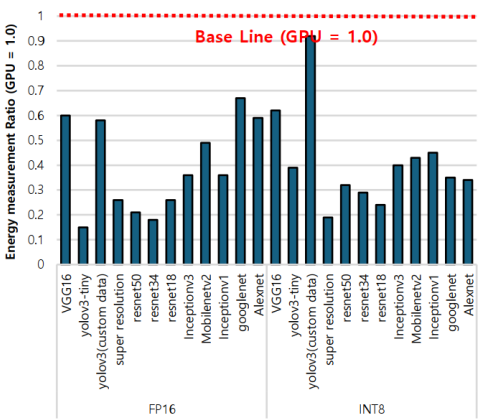

그림 11 DLA 에너지 소모 (GPU대비)

Fig. 11 Energy Consumption by DLA (Normalized to GPU)

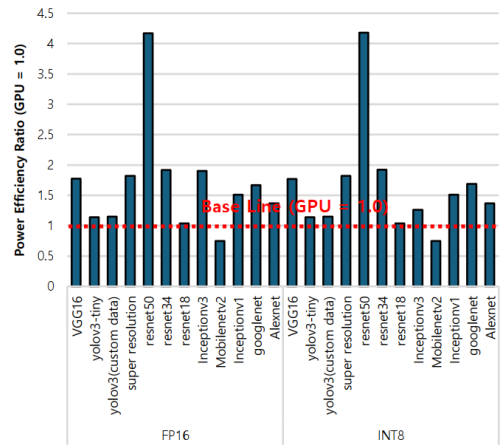

그림 12 DLA 전력 당 성능 (GPU 대비)

Fig. 12 Performance/Power by DLA (Normalized to GPU)

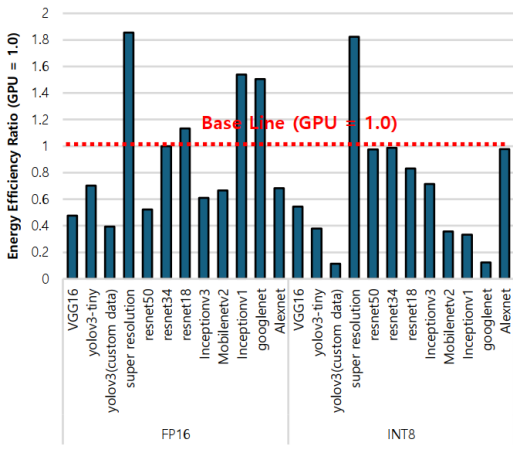

그림 13 DLA 에너지 효율 (GPU 대비)

Fig. 13 Performance/Energy by DLA (Normalized to GPU)

면 VGG16, Inception, Yolo v3 같은 대규모 연산이 필요한 모델에서는 GPU가 더 높은 에너지 효율을 보였다. 이는 DLA가 CNN 모델에 최적화된 가속기임에도 불구하고 연산이 복잡하거나 규모가 큰 경우 GPU가 더 효율적인 결과를 보여준다는 점에서 모델 구조에 따른 효율적인 가속기가 다르다는 것을 보여준다(그림 13).

6. 결 론

본 연구는 임베디드 시스템인 Jetson AGX Xavier의 GPU와 DLA에서 CNN 모델을 실행한 후, 두 가속기의 워크로드 비교 분석을 수행했다. 실험결과, DLA는 CNN 모델에 최적화된 가속기로 GPU에 비해 뛰어난 전력 효율성을 가지며 임베디드 환경에서 CNN 기반 어플리케이션에 적합하지만 대규모 데이터를 처리하거나 여러 모델을 동시에 실행하는 할 경우, 높은 처리량(throughput)과 낮은 지연시간(latency)을 가지는 GPU가 더 적합하다.

DLA는 제한적인 메모리 대역폭으로 인해 여러 모델을 동시에 실행하거나 복잡한 작업을 병렬 처리할 경우 Latency가 증가할 가능성이 있지만, GPU는 다양한 워크로드를 효율적으로 처리하는 구조적인 특징(높은 메모리 대역폭, 병렬 처리능력 등)을 가지고 있어 높은 처리량을 요구하는 환경에서 우수한 성능을 보인다.

따라서 여러 CNN 모델을 동시에 실행하는 환경에서는 DLA의 전력 효율성과 GPU의 높은 처리량, 낮은 지연시간을 고려해 상호 보완적으로 활용하는 것이 효율적이다.

결론적으로, 임베디드 시스템에서 특정 한 가지 모델을 실행하는 상황에서 전체 시스템의 안정적인 운용과 전력 효율성이 높은 DLA가 적합하며, 여러 모델을 동시에 실행하거나 높은 처리량이 요구되는 상황에서는 GPU가 유리하다. 향후 임베디드 시스템에서의 AI 가속기 선택을 위한 추가적인 정보 제공을 위해 다양한 모델과 상황을 고려한 실험을 진행하고 GPU와 DLA의 장단점을 더 명확히 규명해 최적의 가속기 선택 기준을 제시할 수 있을 것이다.

References

[1] Jetson AGX Xavier Series Data Sheet DS-09654-002_v1.8, Copyright © 2014-2022 NVIDIA Corporation.

[2] D. Ghimire, D. Kil, and S.-H. Kim, “A Survey on Efficient Convolutional Neural Networks and Hardware Acceleration,” *Journal of Electronics*, Vol. 11, No. 8, pp. 1745 - 1756, Aug. 2022.

[3] Deep Learning Performance Documentation, URL: https://docs.nvidia.com/deeplearning/performance/dl-performace-convolutional/indx.html (Accessed: Nov. 7, 2024).

[4] NVIDIA TensorRT, URL: https://developer.nvidia.com/tensorrt (Accessed:Nov.7,2024).

[5] M. Capra, B. Bussolino, A. Marchisio, G. Masera, M. Martina, and M. Shafique, “Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead,” *IEEE Access*, Vol. 8, pp. 225134 - 225180, 2020.

[6] Jetson AGX Xavier, URL: http://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-agx-xavier (Accessed: Nov. 7, 2024).

[7] NVIDIA Jetson Xavier AGX System-on-Module, NVIDIA Corporation.

[8] G. Akkad, A. Mansour, and E. Inaty, “Embedded Deep Learning Accelerators: A Survey on Recent Advances,” *IEEE Trans. Artif. Intell.*, Vol. 5, No. 5, pp. 976 - 989, May 2024.

[9] B. Aslan and A. Yilmazer-Metin, “A Study on Power and Energy Measurement of NVIDIA Jetson Embedded GPUs Using Built-in Sensor,” *Proc. IEEE UBMK,* 2022, pp. 112-119.

[10] A. N. Mazumder et al., “A survey on the optimization of neural network accelerators for micro-AI on-device inference,” *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, Vol. 11, No. 4, pp. 532 - 547, Dec. 2021.

[11] NVIDIA TensorRT Developer Guide, URL: https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html (Accessed: Nov. 7, 2024).

권 도 원

2009~2013 해군사관학교 해양학과(학사)
2023~2025 고려대학교 컴퓨터학과(석사)
관심분야는 임베디드시스템, 전력효율

구 건 재

정보과학회논문지
제 52 권 제 2 호 참조