유도무기 AI 연산 가속을 위한 DPU 기반 FPGA 구조 설계 및 적용 방안

송제강, 김종한*
LIG 넥스원

jegang.song@lignex1.com, *jonghan.kim@lignex1.com

Design and Implementation of FPGA-Based DPU Structure for Accelerated AI Inference in Missile Systems

Jegang Song, Jonghan Kim*
LIG Nex1

요 약

본 논문은 유도무기 시스템에서의 AI 알고리즘 적용을 위해, Xilinx 의 Deep Learning Processor Unit(DPU)을 활용한 FPGA 연산 구조를 설계하였다. AI 모델을 Vitis AI 툴체인을 통해 DPU 에 최적화하는 방법을 제시하며, 기존 코어 기반 연산 대비 약 3.5 배 향상되었으며, 모델 정확도는 유지하면서도 연산 효율을 높이는 결과를 확인하였다.

Ⅰ. 서 론

기존 유도무기 시스템에서의 표적 탐지 및 영상 인식 알고리즘은 실시간 처리를 위해 DSP 나 Zynq 의 APU 등 다양한 연산 코어를 복합적으로 활용하는 구조로 설계해왔다. 그러나 이러한 구조는 연산 자원의 과도한 점유와 전력 소모 증가로 인해, 소형화된 유도무기나 제한된 환경에서의 적용에 한계가 존재한다. 이러한 하드웨어 크기 및 소비전력 측면의 문제에 따라 최적화된 새로운 연산 구조의 필요성이 대두되고 있다.

그 중 최근에는 Hailo NPU 등 AI 가속 프로세서를 활용한 AI 알고리즘이 대체 방안이 제시되고 있으며, 여러 HW 가속 방안 중 우주 탐사 시스템에서 적용된[1] FPGA 내부의 Xilinx DPU(Deep Learning Processing Unit) IP 를 이용한 AI 알고리즘 연산을 제안하고자 한다.

본 연구에서는 KV260 개발 보드를 활용하여 DPU IP 를 이용한 AI 모델의 적용 방안과 기존 모델 비교 정확성, 효율성 등을 연구하였다. 이를 통하여 기존 유도무기 시스템의 기존 FPGA 펌웨어 내 일부 블록만 수정하여 적용하는 방법을 제시한다.

Ⅱ. 본론

2.1 DPU IP 를 이용한 펌웨어 설계

본 연구에서는 Xilinx 의 Deep Learning Processor Unit (DPU) IP 를 기반으로 한 펌웨어 설계 구조를 구현하였다. DPU 는 FPGA 내부에서 CNN 기반 모델의 연산을 하드웨어 병렬 구조로 처리할 수 있도록 설계된 AI 전용 가속기로, 기존의 Zynq Ultrascale+ 시리즈 제품군 및 Kria K26 를 통한 소형화, 최신 Versal SoC 제품군에서도 호환 가능하다.

설계에 사용된 하드웨어 플랫폼은 KV260 Vision AI Starter Kit 이며, Block Design 기반으로 DPU Core 를 AXI 인터페이스를 통해 Zynq PS 와 연동되도록 구성하였다. 주요 블록 구성은 아래와 같다.

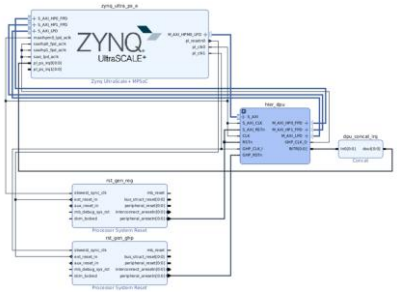

Figure 1. DPU IP 를 이용한 Block Design

2.2 Vitis-AI 를 이용한 모델 최적화

본 연구에서는 YOLOv5n 모델을 DPU 에서 실시간 추론 가능하도록 최적화하기 위해 Xilinx 의 Vitis AI 툴체인을 활용하였다. Vitis AI 는 모델의 Quantization, Compilation, 구조 수정 등을 지원하여 DPU 호환 형식으로 변환할 수 있다.

YOLOv5n 은 PyTorch 기반의 경량 모델로, 기본적으로 SiLU 활성화 함수를 사용하지만 이는 DPU 에서 지원되지 않는다. 이에 따라 Vitis AI Inspector 를 통해 미지원 연산자를 분석하고, 활성화 함수를 DPU 호환 LeakyReLU 로 수정하였다. 이후 PyTorch 모델을 ONNX 로 변환하고, Vitis AI 의 양자화 및 컴파일 과정을 통해 DPU 실행용 .xmodel 파일로 생성하였다.

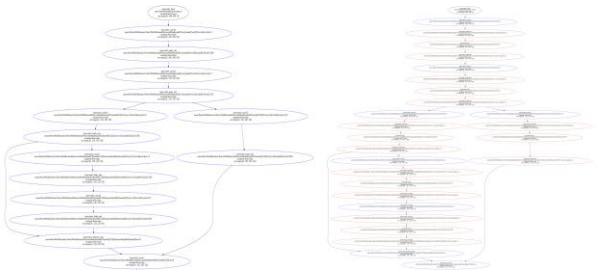

Figure 2. Modified 모델[좌]와 기존 Yolov5n[우]를 Vitis-AI Inspector 를 이용한 활성화 영역 비교

2.2 Vitis-AI 를 이용한 모델 최적화

DPU IP 에 최적화된 모델의 실효성을 평가하기 위해 기존 YOLOv5n 모델과 구조 변경된 DPU 호환 모델의 객체 탐지 성능을 비교하였다. 그 결과, 구조가 변경된 모델은 기존 모델 대비 1.5% 이내의 정확도 차이를 보이며 거의 유사한 수준의 탐지 성능을 유지하였다.

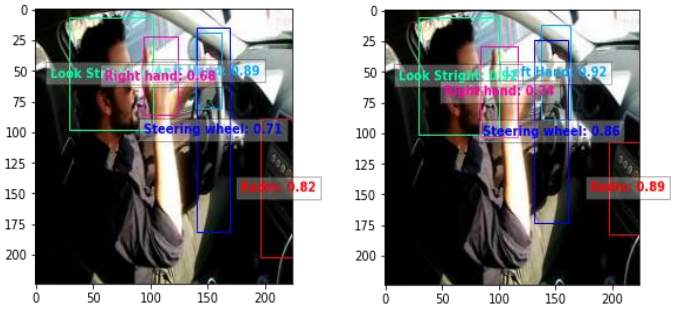

Figure 3. Modified 모델[좌]와 Yolov5n[우]의 정확도 비교

CPU 및 DPU 환경에서 각각 추론 시간을 평가는 아래와 같다. ARM Cortex-A53 CPU 기반에서 수행한 결과는 약 **240ms** 였으며, 동일 입력에 대해 DPU 기반 추론 시 약 **68ms** 로 **3.5 배** 이상 속도 개선을 확인할 수 있었다. 이는 DPU 가 CNN 연산에 특화된 병렬 처리 구조를 바탕으로, 제한된 연산 자원에서도 고속 추론이 가능함을 보여준다.

Ⅲ. 결론

본 연구에서는 유도무기 시스템의 연산 구조를 경량화하고 탐색 성능을 확보하기 위해, Xilinx 의 DPU(Deep Learning Processor Unit) IP 를 활용한 FPGA 기반 AI 연산 구조를 설계하였다. 기존의 DSP 및 APU 기반 알고리즘 대비 연산 효율성과 전력 최적화 측면에서 유리한 구조를 구현하였으며, Vitis AI 툴체인을 활용하여 모델을 DPU 환경에 맞게 구조적으로 간단하게 최적화하여 가속 효율을 높힐 수 있다.

제안된 방식은 기존 FPGA 펌웨어 내 일부 블록만 추가하여 적용할 수 있어, 기사용되던 Zynq Ultrascale+ 제품군뿐만 아니라 Kria K26 기반의 소형 플랫폼 및 Versal SoC 의 AIE 에도 적용 가능하다. 결과적으로, 본 연구는 DPU IP 를 이용한 연산 구조가 유도무기 시스템의 실시간성, 소형화, 전력 최적화 요구를 동시에 만족시킬 수 있음을 보였으며, 향후 국방 임베디드 시스템 내 AI 응용 기술의 적용 가능성을 기술적으로 입증하는 기반을 마련하였다.

ACKNOWLEDGMENT

본 연구는 LIG 넥스원 미사일시스템핵심기술연구소의 지원을 받아 이루어졌음에 감사드립니다.

참 고 문 헌

[1] K. Cosmas and A. Kenichi, "Utilization of FPGA for Onboard Inference of Landmark Localization in CNN-Based Spacecraft Pose Estimation," Aerospace, vol. 7, no. 11, pp. 1- 18, 2020.

[2] 김재명, 강진구, 김용우, “Xilinx DPU 를 이용한 CNN 기반 초해상화 하드웨어 구현,” 대한전자공학회 하계종합학술대회 논문집, pp. 1025-1027, 2022.