2023 IEEE VLSI Review

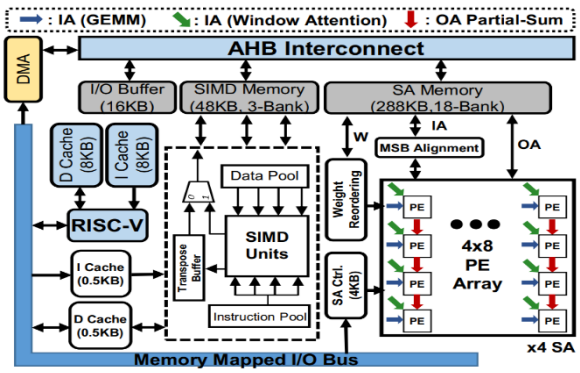
포항공과대학교 전자전기공학과 박사과정 변영훈

Session 16 Advanced NNs

이번 2023 IEEE VLSI의 Session 16은 Advanced NNs라는 주제로 총 5편의 논문이 발표되었다. 이 세션에서는 다양한 분야에서 딥 러닝을 활용하는 프로세서들을 다루는데, 환경에 따라 dynamic한 cost-performance trade-off를 제공하는 방식이 눈길을 끌었다. 본 글에서는 그 중 4개의 논문을 선정해 리뷰해 보았다.

#16-1 A 28 nm 66.8 TOPS/W Sparsity-Aware Dynamic-Precision Deep-Learning Processor

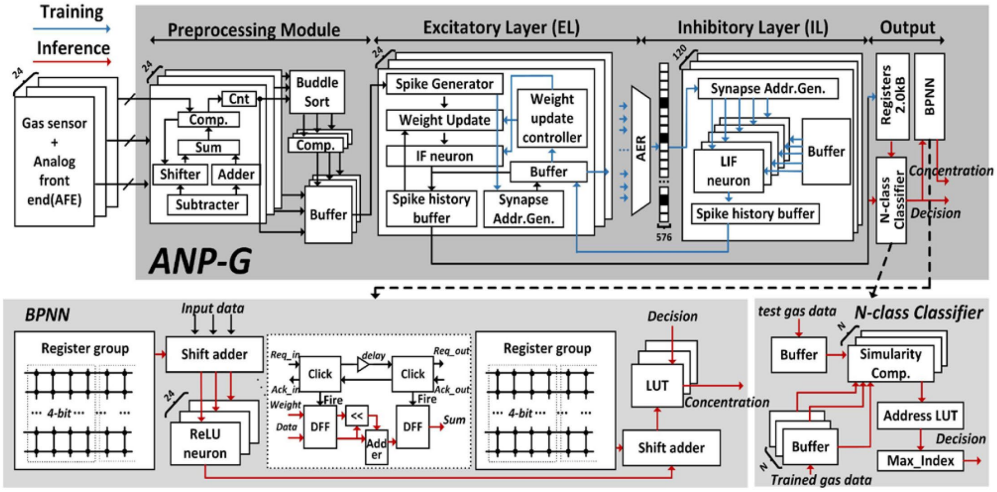
본 논문은 dynamic precision, sparsity aware, heterogeneous DNN accelerator 구조와 이를 기반으로 제작한 ASIC의 성능을 대해 다루고 있다. 그림 1을 보면 해당 architecture는 크게 data reuse가 많은 연산들을 처리하는 systolic array (SA)와 이를 제외한 연산들을 처리하는 SIMD core, 그리고 전체 시스템을 컨트롤하는 RISC-V 코어 영역으로 나누어진다. 해당 아키텍처는 먼저 입력으로 들어온 8-bit 2's complement를 5-bit 1's complement로 변환시킨 뒤, SA에 있는 4-bit multiplier 및 adder tree를 이용해 연산하는 방식을 통해 5b 부터 8b까지 dynamic precision을 지원한다. 또한 sparse 입력의 경우 weight reordering module을 통해 runtime zero gating을 하는데, 이를 통해 sparsity에 따라 energy efficiency의 경우 1.3x - 9.8x, performance의 경우 2.9x – 8x까지 성능을 개선할 수 있다. 해당 구조는 Transformer 뿐만 아니라 MBConv까지 지원하며, 480MHz에서 최대 4.18 TOPS의 성능을 보였다.



[그림 1] 제안하는 processor의 전체 architecture

#16-2 ANP-G: A 28nm 1.04pJ/SOP Sub-mm2 Spiking and Back-propagation Hybrid Neural Network Asynchronous Olfactory Processor Enabling Few-shot Class-incremental On-chip Learning

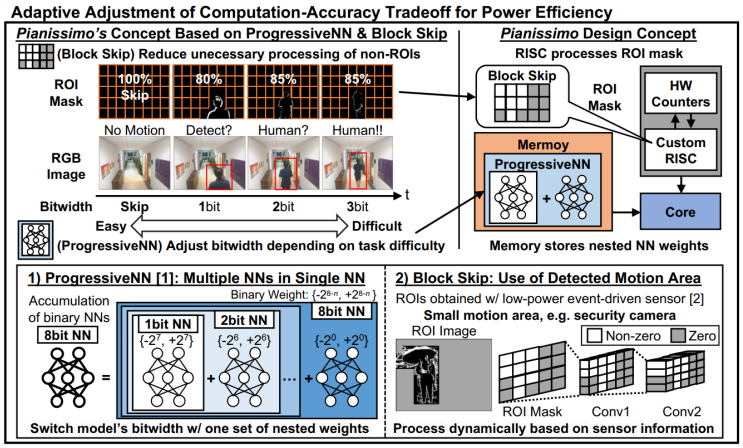
본 논문에서는 가스 인식, 농도 추정, 그리고 incremental learning을 지원하는 Spiking Neural Network기반 프로세서 ANP-G를 제안한다. 일반적인 gas sensor의 경우 시간이 지날수록 sensitivity가 변하는 특징이 있는데, 이로 인한 가스 인식 및 농도 추정의 정확도 저하를 해결하기 위해서는 retraining이나 calibration을 필요로 한다. 이를 위해 ANP-G는 인식을 담당하는 two-layer SNN과 학습을 담당하는 three-layer BPNN으로 구성되었다. 제안하는 프로세서는 가스 센서의 drift로 인한 정확도 저하를 극복하고, 가스 관련 작업에서 효과적인 성능을 보여주며 previous work[1]과 비교해 0.55V에서 4.09x 더 높은 pJ/SOP를 달성하였다.



[그림 2] ANP-G의 전체 architecture

#16-4 Pianissimo: A Sub-mW Class DNN Accelerator with Progressive Bit-by-Bit Datapath Architecture for Adaptive Inference at Edge

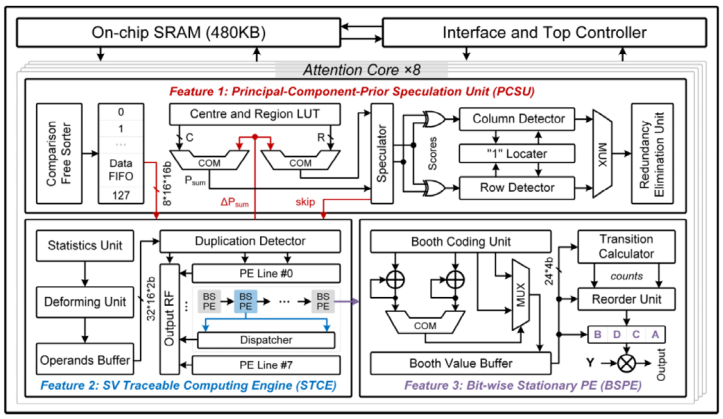
Pianissimo는 RISC 코어를 사용해 상황에 따라 bit-by-bit으로 모델의 성능을 조정하는, edge device에서 저 전력으로 동작할 수 있는 class inference accelerator이다. 논문에서는 크게 두 가지 주요 기능을 소개하고 있다. 첫 번째는 progressiveNN이라는 기법으로, memory에 모델을 저장할 때 bit-by-bit으로 저장한 뒤, inference 난이도에 따라 load하는 모델의 bit을 조절하는 방식을 사용한다. 이 방식을 사용하면 하나의 모델만 저장해도 성능과 cost간의 trade-off를 지원할 수 있다는 장점이 있다. 다음으로 Block Skip이라는 방식에서는 RISC에서 moving object들을 region of interest (ROI)로 두고, 해당 범위를 제외한 부분들에 대한 연산을 생략한다. 40nm공정에서 제작된 이 칩은 mobileNetV1 사용 시 0.7V에서 793-1032 uW를 소비하며 0.49-1.25TOPS/W 를 달성하였다.



[그림 3] ProgressiveNN과 Block Skip을 사용한 Pianissmo의 전체 컨셉 설명

#16-5 A 28nm 77.35TOPS/W Similar Vectors Traceable Transformer Processor with Principal-Component-Prior Speculating and Dynamic Bit-wise Stationary Computing

본 논문은 Transformer architecture 의 quantized attention 연산에서 발생하는 similarity 에 기반해 redundant 한 연산을 제거하는 방식을 통해 energy efficient 하게 동작하는 Transformer process 를 제안한다. 해당 프로세서는 크게 Principal-Component-prior Speculation Unit (PCSU), Similar vector Tracked Computing Engine (STCE), Bit-wise Stationary Processing Element (BSPE) 모듈로 구성되어 있다. 먼저 PCSU 에서는 attention 을 계산하기 전 $|Q \times K^T|$ 연산에서 magnitude 가 큰 경우만 시행하고 작은 경우들은 생략하는 방식으로 중복 계산의 28.4%를 제거한다. 다음으로 STCE 에서는 similar vector 들에 대한 연산을 병렬화 함으로 써 곱셈의 42.2%를 줄인다. 마지막으로 BSPE 모듈에서는 booth value (BV)를 사용하는데, transition energy 를 줄이기 위해 reordering 을 활용하였고, 곱셈 연산에 드는 에너지를 1.47 배 줄였다. 최종적으로 제안하는 프로세서는 소모 에너지를 2.81 배 줄이고 동작 속도를 3.71 배 올렸으며 77.35 TOPS/W 를 달성하였다.



[그림 4] 제안하는 프로세서의 전체 아키텍처

참고문헌

[1] C. Frenkel and G. Indiveri, "ReckOn: A 28nm Sub-mm2 Task-Agnostic Spiking Recurrent Neural Network Processor Enabling On-Chip Learning over Second-Long Timescales," 2022 IEEE International Solid-State Circuits Conference (ISSCC), Feb. 2022

저자정보

변영훈 박사과정 대학원생

- 소속 : 전자전기공학과
- 연구분야 : Deep learning model compression
- 이메일 : byh1321@postech.ac.kr
- 홈페이지 : sites.google.com/view/epiclab/member/yhbyun