2023 IEEE ASSCC Review

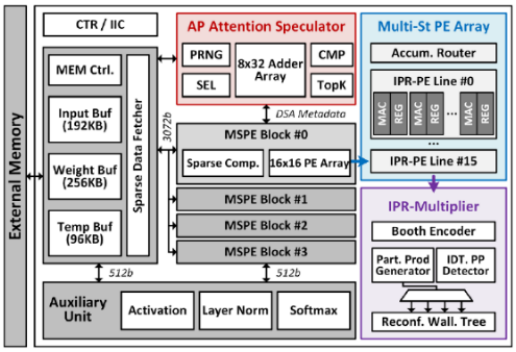
한양대학교 신소재공학과 박사과정 송충석

Session 9. Energy-Efficient Application-Specific Processors

이번 2023 IEEE ASSCC의 Session 9는 Energy-Efficient Application-Specific Processors라는 주제로 총 4편의 논문이 발표되었다. 일반적인 연산 및 데이터 처리를 할 수 있는 general purpose 하드웨어(CPU 혹은 GPU)와 달리 특정 task에 특화된 설계를 할 수 있는 주문형 반도체(ASIC)의 특성에 맞게 하드웨어적 최적화를 진행하여 에너지 소모를 줄이는 프로세서를 중국에서 3편, 미국에서 1편 발표하였다. 각 논문들은 트랜스포머(transformer)(9-1), voice activity detection(9-2), CRYSTALS-KYBER(9-3), SNN 학습에 최적화된 프로세서(9-4)를 각각 발표했고, 본 리뷰에서는 9-1, 9-2, 9-4를 다루어 보고자 한다.

#9-1 논문은 자연어 처리 분야를 포함한 다양한 분야에서 각광받고 있는 트랜스포머 알고리즘을 효율적으로 연산할 수 있는 프로세서를 발표하였다. 본 논문의 프로세서는 크게 3가지의 특징을 가지는데 : (i) throughput과 에너지 소모를 개선하기 위해 dynamic sparse attention(DSA) 방식을 이용하였다. 인풋 토큰으로부터 생성된 Q(query), K(key), V(value)로부터 얻은 attention value에 sparsity를 적용하는 것으로 임의의 난수 마스크를 생성하여 sparsity를 구현하였고, (ii) 정해진 하나의 data 유형만(input, weight, output) 레지스터에 저장(X-stationary)하여 연산을 하는 기존 프로세서에 비해 본 프로세서는 matrix 크기에 맞게 다양한 데이터를 유기적으로 레지스터에 저장할 수 있는 데이터 흐름도를 개발 및 적용하여 연산 효율성을 증가시켰고, (iii) 마지막으로 softmax연산에서 모든 인풋에 같은 값을 더하거나 빼도 softmax 결과가 바뀌지 않는다는 것을 이용하여 attention value를 구하기 위한 softmax 연산 시 $Q \cdot K^T$(softmax 연산의 인풋)의 결과를 모두 계산하여 softmax를 취하는 것이 아닌, 각각의 $Q \cdot K^T$의 차이만을 계산하여 중복되는 계산 결과를 생략하여 연산 효율성을 증가시켰다.
첫번째 특징으로 인해 정확도는 오직 1퍼센트정도만 감소했음에도 불구하고 Q, K, V 연산은 34.6%, attention 연산은 89.5% 줄일 수 있었고, 두번째 특징으로 인해 평균적으로 21.9%의 에너지를 절약, 마지막 세번째 특징으로 전력소모를 1.91배 감소시켰다. 본 프로세서는 28nm 1P8M CMOS 공정으로 제작되었고 sparsity 90퍼센트를 가지는 INT8연산에서 6.53TOPS (Trillion Operation Per Second), 49.7TOPS/W (TOPS per Watt)의 성능을 달성

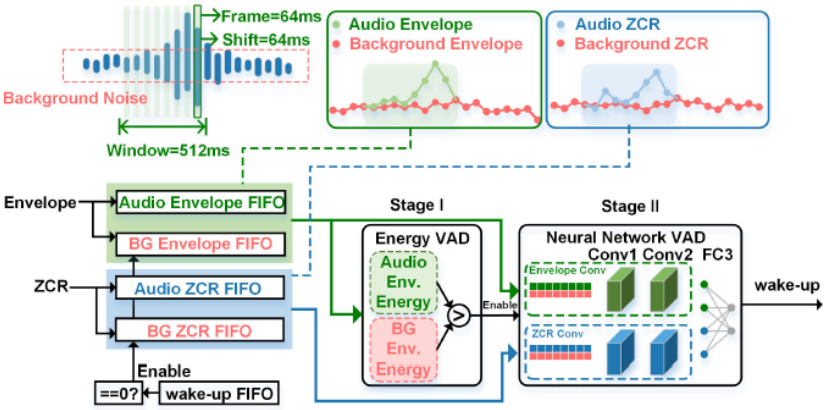
했다.



[그림 1] #9.1에서 제안한 프로세서의 전체 구조

#9-2 논문은 IoT 장치에 필요한 노이즈에 강한 Voice Activity Detection(VAD) 프로세서를 발표하였다. VAD는 타겟 시스템에서 소리를 탐지하기 위해 항상 디바이스가 켜져야 하는 always-on 디바이스 특성상 저전력으로 구현되어야 하며 높은 노이즈에도 정상적으로 작동할 수 있어야 한다. 따라서 본 논문에서는 포락선(envelope)과 zero-crossing rate(ZCR)을 입력으로 받아 인공 신경망 분류기(classifier)를 통해 소리를 구분하였다. 인공 신경망 분류기는 2개의 1차원 convolution layer와 wake-up 신호를 만들기 위한 fully connected layer로 이루어져 있다. 512ms의 길이를 가지는 소리 신호를 64ms의 프레임으로 나누어 입력을 받은 후, 오디오의 특징은 매 프레임마다 업데이트를, 배경(background, noise)의 특징은 소리신호가 감지되지 않을 때(wake-up signal)마다 업데이트를 진행하였다.

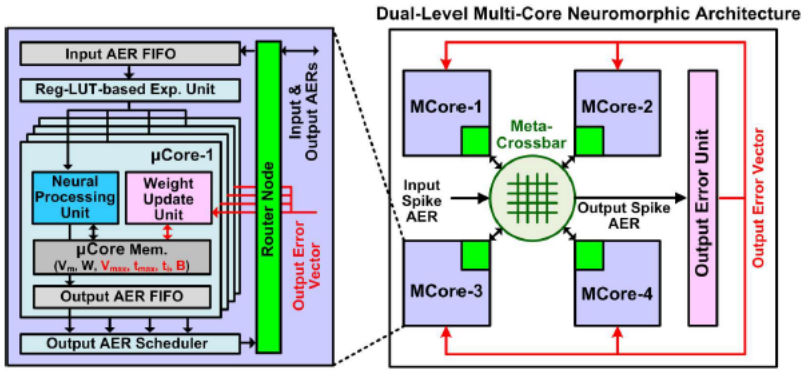
본 프로세서는 65nm CMOS 공정으로 0.07mm^2의 면적으로 제작되었고, 10dB babble 노이즈에 대해 93.56%, -5dB Volvo 노이즈에 대해 91.53%의 준수한 성능을 보여주었다. 특히, 0.4V의 동작전압에서 28nW의 저전력 동작을 구현하였다는 것이 눈에 띄었다.



[그림 2] #9.2에서 제안한 VAD 시스템

#9-4 논문은 인간의 뇌의 메커니즘과 가장 유사하다고 평가받는 Spiking Neural Network(SNN)에 최적화된 뉴로모픽 프로세서를 발표했다. 특히, SNN은 낮은 에너지 소모와 낮은 latency를 요구하는 어플리케이션들에 적합한데, 아직까지 많은 뉴로모픽 프로세서들은 이러한 SNN을 높은 수준으로 구현하기에는 한계가 존재했다. 본 논문에서는 multi layer fully connected(FC) SNN을 빠르게 연산할 수 있도록 구현함과 동시에 프로세서 내부에서 학습까지 가능케 하였다. 하나의 매크로 코어(MCore)에 FC layer의 일부 혹은 전부를 맵핑하고, 마이크로 코어(uCore)에 다수의 LIF 스파이킹 뉴런을 이용하였다. 본 프로세서는 4가지 특징을 가지는데 : (i) 모든 MCore들은 이벤트기반으로 서로 독립적으로 동작하고(feedforward), 모든 uCore들은 동기화되어 학습을 하여(backward) hybrid parallelism을 프로세서에 구현해 throughput을 향상시켰고, (ii) 스파이크 전달 시 meta-crossbar 라우팅 방식을 제안하여 라우팅 오버헤드를 완화시켰고, (iii) 에너지 소모를 줄이기 위해 각 뉴런들은 최대 한번만 활성화되고, 인풋 이미지 픽셀을 하나의 스파이크로 변환시켰다. (iV) 마지막으로 본 연구팀에서 사전에 제안한 DeepTempo 학습법을 칩에 적용시켜 17.6퍼센트의 면적을 줄였다.

본 프로세서는 65nm CMOS 공정을 이용하여 7.2mm^2 면적으로 제작되었으며 MNIST 데이터셋을 이용해 96.06퍼센트의 정확도를 달성했다. 학습/추론에서 각각 802/2270 frame/s, 10.32/3.04 pJ/SOP(synaptic operation) 의 성능을 달성하였다. 학습과정을 온 칩에 구현하기 상대적으로 어려운 DNN에 비해 SNN의 장점을 살려 온 칩에 학습까지 구현한 점이 인상깊다. 다만, 좀 더 거대규모의 network 혹은 복잡한 데이터셋에 대한 검증은 추가로 필요해 보인다.



[그림 3] #9.4에서 제안한 뉴로모픽 프로세서 구조

저자정보

송충석 박사과정 대학원생

- 소속 : 한양대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : scs940430@naver.com
- 홈페이지 : https://sites.google.com/site/dsjeonglab1/home