채널 병렬화 방법을 통한 CNN 가속기 PE Utilization 향상

장건희, 김형원*

충북대학교, *충북대학교,

geonhuijang@chungbuk.ac.kr, *hwkim@chungbuk.ac.kr,

Enhancing PE Utilization in CNN Accelerators Through Channel Parallelization

Geonhui Jang, Hyungwon Kim*

Chungbuk National Univ., *Chungbuk National Univ

요 약

본 논문에서는 3x3 PE Array 구조에서 1x1 컨벌루젼 연산 시 발생하는 PE Utilization 저하 문제를 해결하기 위한 방법을 제안한다. 이를 개선하기 위해 두 가지 병렬화 방법을 검토하였으며, 그 중 채널 병렬화 방법이 PE Utilization을 높이는 데 더 효과적임을 확인했다. 또한, 이 방법을 U-Net의 Transposed 컨벌루젼에도 적용할 수 있음을 확인했다. 제안된 기법을 사용하면 기존 방식에 비해 PE Utilization을 11.11%에서 87.78%로 향상됨을 확인했다.

Ⅰ. 서 론

컨벌루젼 신경망(CNN)은 이미지 분류, 객체 검출, 세분화 등 다양한 컴퓨터 비전 분야에서 비약적인 발전을 이루어왔다. 특히 VGG16, YOLO, U-Net 등과 같은 다양한 CNN 모델들에서 3x3 컨벌루젼 연산은 전체 연산의 대부분을 차지하며, 이미지의 중요한 특징을 추출하는 핵심적인 역할을 한다. 이에 따라, 많은 CNN 하드웨어 가속기들은 3x3 또는 3x1 PE 구조를 사용해 이러한 3x3 컨벌루젼 연산을 가속화하고 있다. 이러한 구조는 높은 연산 성능을 제공하며, 다수의 병렬 연산을 효율적으로 처리할 수 있어 3x3 컨벌루젼에서 높은 PE Utilization을 달성할 수 있다. 그러나 1x1 컨벌루젼의 경우, 커널 크기가 작아져서 PE Utilization이 낮아지는 문제가 발생한다. 이에 본 논문에서는 채널 병렬화 기법을 통해 1x1 컨벌루젼에서의 PE Utilization을 개선하는 방법을 제안하며, 나아가 U-Net의 2x2 stride 2 Transposed 컨벌루젼에서도 동일한 방식으로 PE Utilization을 향상시키는 방법을 제안한다.

Ⅱ. 본론

A. 1x1 컨벌루젼

3x3 구조를 사용하는 PE Array에서 3x3 컨벌루젼을 수행할 때는 하나의 PE Array가 한 채널의 컨벌루젼 부분합을 계산한 후, Adder Tree를 통해 여러 채널의 결과를 합산하여 최종 컨벌루젼 결과를 얻는다. 그러나 1x1 컨벌루젼을 수행할 경우, 9개의 PE 중 단 하나만 컨벌루젼 연산에 활용되므로 PE Utilization이 낮아지는 문제가 발생한다.

이 문제를 해결할 수 있는 방법은 두 가지 병렬화 방법이 있다. 첫 번째는 픽셀 병렬화 방법이다. 이 방법에서는 9개의 PE를 각각 다른 픽셀에 대해 연산하도록 하여 모든 PE를 활용하는 방식이다. 이를 통해 PE Utilization을 높일 수 있지만, 이 경우 채널 간 합산을 위한 Adder Tree가 9배 더 필요하고, 매 사이클마다 9개의 출력값이 생성되므로 이를 저장하기 위해 Bandwidth가 9배 큰 메모리를 요구하기 때문에 높은 리소스 오버헤드가 발생한다는 단점이 있다.

두 번째 방법은 채널 병렬화 방법이다. 제안하는 채널 병렬화 방법은 그림 1과 같이 각 PE가 서로 다른 채널에 대해 독립적으로 컨벌루젼 연산을 수행하고, PE 내의 adder를 채널간에 합산에 활용하여 PE Array가 Adder Tree의 역할을 하도록 하는 재구성하는 방식이다. 이 구조를 통해 PE Array에서 8개 채널의 합산 결과를 얻을 수 있으며, 최종적으로 Adder Tree는 32개 채널의 합산 결과를 생성하게 된다. 이 방법은 추가적인 Adder Tree를 요구하지 않으며 낮은 On-chip Bandwidth를 유지할 수 있다는 장점이 있다.

따라서 제안하는 채널 병렬화 방법이 기존 픽셀 병렬화 방법보다 PE Utilization을 높이는데 더 유리하며, 대부분 CNN 모델의 채널 수가 9의 배수가 아닌 8로 나누어 떨어지기 때문에 9개의 PE 대신 8개의 PE를 사용하는 것이 적합하다.

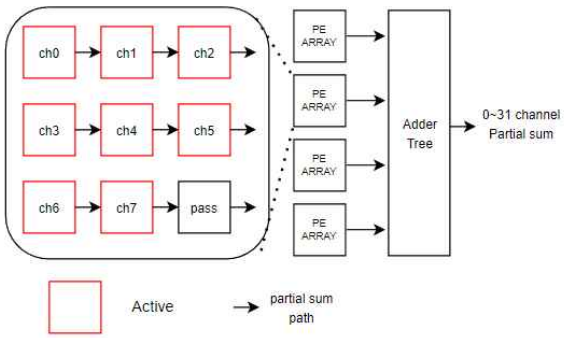

그림 1. 제안하는 1x1 컨벌루젼 병렬화 방법.

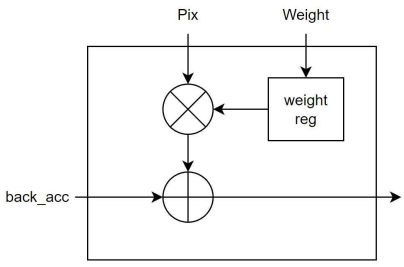

그림 2. PE 구조.

B. Transposed 컨벌루션

U-Net은 주로 의료 영상 분할에 사용되는 모델로, 2x2 stride 2 Transposed 컨벌루션을 포함한다. Transposed 컨벌루션은 세그멘테이션 모델에서 자주 사용되며, 인코더에서 축소된 이미지를 디코더를 통해 원래 크기로 복원하는 데 활용된다. 그림 3은 2x2 stride 2 Transposed 컨벌루션 연산 과정을 설명하며, 하나의 픽셀이 4개의 가중치와 곱해져 출력이 2x2로 확장되는 과정을 보여준다. 이 과정은 1x1 컨벌루션과 유사하게 진행되며, 하나의 출력에 대해 하나의 픽셀과 가중치가 곱해지며, Adder Tree를 통해 여러 채널의 결과를 합산하여 최종 컨벌루션 결과를 얻는다. 따라서, 그림 1에서 제안된 1x1 컨벌루션 방식을 통해 Transposed 컨벌루션을 구현할 수 있다

그러나 1x1 컨벌루션과 달리, 2x2 stride 2 Transposed 컨벌루션에서는 하나의 픽셀이 4개의 가중치와 연산되므로 픽셀을 재활용하는 과정이 필요하다. 이를 구현하기 위해, 그림 4에서 제시된 바와 같이 PE 내의 레지스터에 weight를 저장하는 weight stationary 방식 대신, 상대적으로 작은 weight를 반복해서 읽고 픽셀을 PE 내의 레지스터에 저장하는 input stationary 방식을 사용하면, 픽셀을 효과적으로 재활용하면서 높은 PE Utilization을 유지하며 Transposed 컨벌루션을 수행할 수 있다.

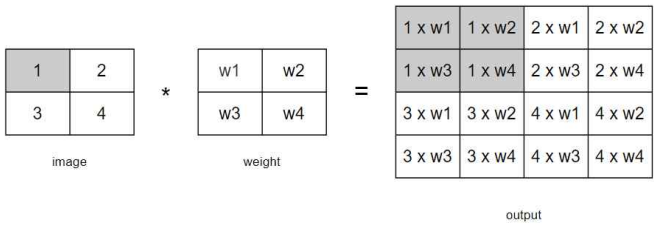

그림 3. 2x2 stride 2 Transposed 컨벌루션 연산 과정

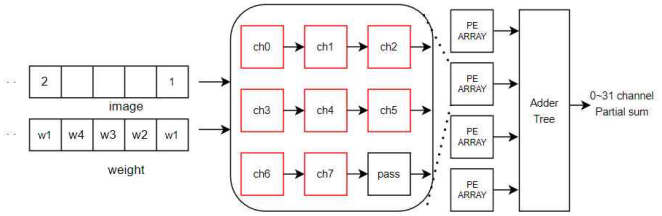

그림 4. 제안하는 2x2 stride 2 Transposed 컨벌루전

Ⅲ. 실험 결과

본 논문에서는 Verilog 언어를 사용하여 RTL 설계를 수행하였으며, 하드웨어 기능 검증을 위해 Vivado 2020.2 툴을 사용하였다. 컨벌루전 모듈의 검증을 위해 Pytorch 프레임워크를 활용하여 테스트 데이터를 생성하였다. 총 PE의 개수는 576개(3x3x4x16)로, 테스트 데이터의 크기는 (1, 32, 100, 100)이다. PE Utilization은 컨벌루전 과정동안 활성화된 PE의 수를 전체 PE의 수로 나누어 계산하였으며, 제안된 방식을 적용하면 PE Utilization이 기존 11.11%에서 88.78%로 증가하는 것을 확인했다.

data 크기	기존 방식의 PE Utilization	제안한 방식의 PE Utilization
(1,32,100,100)	11.11%	88.78%

표 1. PE Utilization 비교

Ⅳ결론

본 논문에서는 3x3 PE Array 구조에서 1x1 컨벌루전의 PE Utilization이 낮아지는 문제를 해결하기 위해 채널 병렬화 방법을 제안한다. 제안된 방법을 통해 단일 버퍼와 추가적인 Adder Tree없이 PE Utilization을 기존 11.11%에서 88.78%까지 향상시킬 수 있음을 확인하였다.

ACKNOWLEDGMENT

This work was supported by Regional Leading Research Center (RLRC) of the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A5A8026986) and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01304, Development of Self-Learnable Mobile Recursive Neural Network Processor Technology). It was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Communication Technology Research Center support program (IITP-2024-2020-0-01462) supervised by the IITP (Institute or Information & communications Technology Planning & Evaluation).

참 고 문 헌

[1] Lee, Dong-Yeong, et al. "High-Speed CNN Accelerator SoC Design Based on a Flexible Diagonal Cyclic Array." Electronics 13.8 (2024): 1564.

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.