

최대 풀링 레이어를 위한 저면적 재구성 가능한 하드웨어 구조

김은서, 신동엽, 임용석
한국전자기술연구원 스마트네트워크센터

golroom97@keti.re.kr, dongyeob.shin@keti.re.kr, busytom@keti.re.kr

Low Area Reconfigurable Hardware Architecture for Max Pooling Layer

Eun-Seo Kim, Dongyeob Shin, Yong-Seok Lim

Smart Network Center
Korea Electronics Technology Institute

요 약

합성곱 신경망(Convolutional Neural Network)의 적용 범위가 확장됨에 따라 풀링 레이어(Pooling Layer)의 필터 크기 및 스트라이드 크기 또한 다양해지고 있다. 이에 본 논문에서는 풀링 레이어의 필터 크기와 스트라이드 크기를 가변적으로 사용할 수 있는 하드웨어 구조를 제안한다. 결과적으로 재구성 가능한 최대 풀링 하드웨어 구조는 최대 풀링 연산을 처리하는 면적을 감소시킬 수 있다.

I. 서 론

합성곱 신경망(Convolutional Neural Network)은 이미지 분류, 객체 인식 분야에서 가장 뛰어난 성능을 보인다. 합성곱 신경망은 특징 추출(Feature Extraction)과 분류(Classification) 기능으로 나누어지며, 이 중 특징 추출은 일반적으로 특징맵(Feature Map) 연산을 위한 합성곱 레이어(Convolutional Layer)와 특징맵의 크기를 줄여 과적합(Overfitting)을 방지하는 풀링 레이어(Pooling Layer)로 구성된다[1]. 풀링의 종류에는 최대 풀링(Max Pooling), 평균 풀링(Average Pooling) 등이 있으며 특히 최대 풀링은 대부분의 합성곱 신경망에서 사용된다. 최대 풀링은 신경망 구조에 따라 적절한 필터(Filter) 크기와 스트라이드(Stride)를 가질 수 있다. 최근 NASNet 과 같이 NAS(Neural Architecture Search)를 통해 자동 생성하는 최적의 신경망 구조들은 2x2, 3x3, 5x5, 7x7 등 다양한 필터 크기 중 최적의 필터 크기를 갖도록 설계되고 있다[2].

본 논문에서는 다양한 필터 크기의 최대 풀링을 지원할 수 있는 저면적 재구성 가능한 하드웨어 구조를 제안한다. 기존의 최대 풀링 하드웨어는 고정된 필터 크기에 적합하게 설계되어 있어, 이를 가변하기 위해서는 큰 면적의 비효율적인 구현이 불가피하다[3]. 제안하는 하드웨어는 폴딩(Folding) 구조를 통해 면적 오버헤드를 최소화하며 필터 크기를 가변할 수 있도록 설계되었다.

II. 본론

2.1. 최대 풀링 레이어

최대 풀링 레이어는 합성곱 레이어의 연산 결과인 특징 맵을 입력으로 하여 필터 크기와 스트라이드에 따라 특징 맵의 크기를 감소시킨다. 그림 1은 2x2 필터, 스트라이드 2인 경우와 3x3 필터, 스트라이드 2인

경우의 최대 풀링 동작을 각각 나타낸다. 그림 1과 같이 최대 풀링은 필터 윈도우 영역의 픽셀 중 가장 큰 값만 추출하여 다음 레이어의 입력이 될 축소된 특징 맵을 생성한다. 하드웨어 구현 시, 필터 크기에 따라 최대 풀링 하드웨어에 입력되는 픽셀 개수 및 비교 연산 횟수가 달라지며, 스트라이드 크기에 따라 처리해야 하는 픽셀들이 달라지는 것을 고려해야 한다.

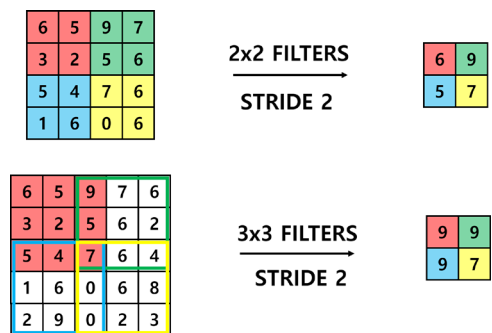


그림 1. 최대 풀링 연산

2.2 제안 기법

먼저 기존 기법과 제안 기법에서 공통적인 최대 풀링 하드웨어의 입력 부분은 다음과 같다. 그림 2는 32x32 입력 특징 맵의 최대 풀링 하드웨어 입력 예시이다. 그림에서는 입력 특징 맵을 열 단위(32x1)로 처리(행 단위도 동일)하는 것을 가정한다. 그림 2처럼 스트라이드의 크기가 2일 때 첫 번째 열인 column[0]에 존재하는 32개의 데이터 중 짝수 번째 행의 데이터가 최대 풀링 연산의 시작점이다. 시작점이 되는 데이터를 기준으로 연속적인 7개의 데이터가 하나의 풀링 연산의 입력이 된다. 그림 2의 최대 풀링 하드웨어 입력 부분은 MUX와 FIFO로 구성되어 최대 풀링 오퍼레이터(Max Pooling Operator) 하드웨어 앞단에 위치한다.

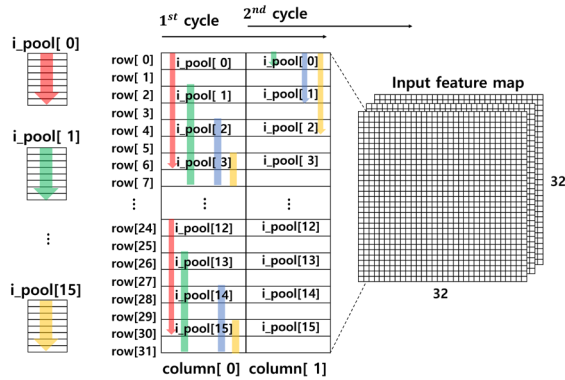


그림 2. 최대 풀링 하드웨어 입력 데이터 예시

그림 3 은 기존 필터 크기를 가변하는 최대 풀링의 하드웨어[3] 구조를 나타낸다. 그림 3 과 같이 기존 최대 풀링 오퍼레이터 하드웨어는 입력 픽셀에 대한 최댓값 비교 연산을 동시에 수행하기 때문에 가변하는 필터 크기 중 가장 큰 필터 크기에 따라 면적이 결정되고, 이는 그보다 작은 필터 크기로 동작 시에 자원을 낭비하는 요소가 된다. 예를 들어, 그림 3 에서와같이 2x2, 3x3 필터를 가변할 수 있도록 구현하는 경우, 2x2 필터 크기로 최대 풀링 시 3x3 최대 풀링에 맞춰 설계된 최댓값 비교 연산기의 하나의 입력은 사용하지 않게 된다. 즉, 기존 최대 풀링 하드웨어는 가변하는 필터의 크기 차이가 벌어질수록 효율이 떨어질 수 있다.

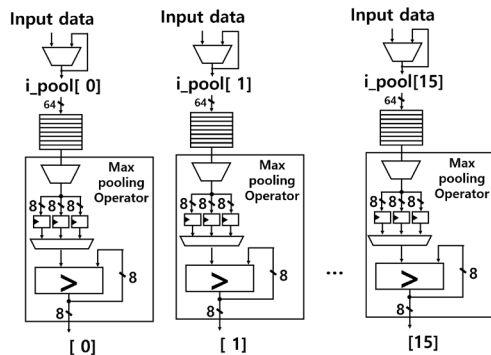


그림 3. 필터 크기를 가변하기 위한 기존 최대 풀링 하드웨어 구조

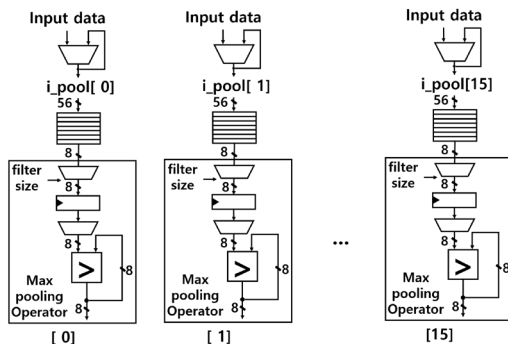


그림 4. 필터 크기를 가변하기 위한 제안 최대 풀링 하드웨어 구조

본 논문에서 제안하는 하드웨어 구조는 그림 4 에 나타나 있다. 그림 4 에서 제안하는 최대 풀링 하드웨어는 풀링 구조를 활용하여 최댓값 비교 연산을 두 개의 피연산자에 대해서만 수행하기 때문에 연산을 처리하는 면적이 필터의 크기에 의해 제한되지 않는다. 또한 1) 필터의 크기에 따라 입력 데이터 순서를

결정하는 MUX 를 통해 비교 연산자의 입력이 될 수 있는 데이터의 개수를 가변하고 2) 비교 연산의 반복 횟수는 카운터를 통해 제어하여, 단일 하드웨어로 필터의 크기를 2x2 부터 7x7 까지 다양하게 사용하는 것이 가능하다. 동시에 제안하는 하드웨어 구조는 비교 연산기의 모든 입력을 사용하여 높은 하드웨어 활용도를 가질 수 있다.

2.3. 구현 결과

표 2 는 기존 최대 풀링 하드웨어 구조와 제안하는 하드웨어 구조 결과를 비교한 것이다. 하드웨어 구현 결과는 합성 결과를 사용하였으며, 합성을 위해서는 시놉시스(Synopsys) 디자인 컴파일러(Design Compiler) 툴과 130nm CMOS 공정 라이브러리를 사용하였다. 10MHz 클록 주파수에서 합성하였고, 기존 하드웨어도 공정한 비교를 위해 동일한 조건에서 합성하여 결과를 비교하였다(redesign). 제안하는 최대 풀링 하드웨어는 9,265um² 의 면적을 가져, 기존 하드웨어의 48,504um² 와 비교하여 80.9%의 면적을 줄일 수 있었다.

	기존[3]	제안
공정 [nm]	130	130
동작 주파수 [MHz]	10	10
면적 [um ²]	48,504	9,265

표 2. 최대 풀링 하드웨어 구현 결과 비교

III. 결론

본 논문에서는 다양한 필터 크기를 갖도록 재구성 가능한 최대 풀링 하드웨어를 저면적으로 구현하고자 하였다. 제안하는 최대 풀링 하드웨어는 기존의 병렬 처리 구조가 아닌 풀딩을 통한 직렬 처리 방식을 통해 하드웨어 면적을 약 80.9% 감소시킬 수 있었다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부의 재원으로 정보통신기획평가원의 정보통신방송기술개발지원사업의 연구결과로 수행되었습니다 (IITP- 2020-0-01077, IoT 다중 인터페이스 기반의 데이터센싱, 엣지컴퓨팅 분석 및 데이터공유 지능형 반도체 기술 개발)

참 고 문 헌

- [1] H. Gholamalinezhad, and H. Khosravi, "Pooling methods in deep neural networks, a review," arXiv, no. September, 2020.
- [2] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le., "Learning transferable architectures for scalable image recognition," arXiv: 1707.07012, 2017. 2, 6
- [3] L. Du et al., "A reconfigurable streaming deep convolutional neural network accelerator for Internet of Things," IEEE Trans. Circuits Syst.I, Reg. Papers, vol. 65, no. 1, pp. 198- 208, Jan. 2018.