목 차

1. 연구제목
2. 연구의 배경/필요성
3. 연구 목적
4. 연구 내용, 범위 및 방법
5. 소요되는 주요 연구시설 및 기기
6. 연구일정
7. 기대되는 성과 및 활용 방안
8. 참고문헌
9. **연구제목**

국문: 뉴로모픽 시스템을 위한 컨볼루션 구조의 병렬화 설계 방법

영문: Synthesis of Activation-Parallel Convolution Structure for Neuromorphic Architecture

1. **연구의 배경/필요성**

최근 컴퓨팅 기술의 발달로 인해, 딥러닝(Deep Learning)을 이용한 인공지능에 대한 관심이 증가하고 있다. 특히, Convolutional Neural Network(이하, CNN)을 이용한 이미지 및 영상 인식 방법에 대해 활발한 연구가 진행되고 있지만, 기존의 폰 노이만 방식의 컴퓨터 구조로는 CNN이 가지고 있는 많은 가중치 파라미터와 메모리와 프로세서 사이의 병목 현상, 또한 그로 기인한 큰 전력 소모가 CNN을 모바일 시스템에 구현하기에 어렵게 하고 있다. 따라서, 기존의 구조에서 구현하기보단, 인간의 뇌와 유사한 방식으로 데이터를 처리하는 뉴로모픽 시스템 방식으로 구현함으로 이것을 해결할 수 있지만, 여전히 컨볼루션 층(Convolution layer)에서의 버스(Bus)의 대역폭(Band-width)의 한계로 인한 처리율(throughput)의 저하를 해결하는 데 어려움이 존재한다. 따라서, 처리율을 높이기 위한 병렬 설계 방법이 요구된다.

1. **연구 목적**

본 연구에서, 시간 지연 신경망(Time-Delay Neural Network, 이하 TDNN)으로 변환시킨 CNN을 뉴로모픽 시스템에 구현하고, CNN의 컨볼루션 층을 병렬화함으로써, 2D 컨볼루션 연산을 빠르게 수행하도록 하는 방법을 제시한다. 따라서, 심층 신경망(Deep Neural Network)을 뉴로모픽 칩(Chip)에 구현할 때, 한정된 공간 안에 처리율을 최대한 증가하도록 하는 설계 방식을 제안한다.

1. **연구 내용, 범위 및 방법**

* 컨볼루션 층의 병렬화 방법
  + 컨볼루션 커널(Kernel)의 Unrolling 방법 개발
  + 2D 컨볼루션의 stride에 따라 가능한 최적화 방법 제시
* 입력 이미지 및 영상에 대한 전/후 처리 문제
  + Border handling: 입력 영상의 2D 컨볼루션 연산에 대한 가장자리 영역 처리 방법
  + Alignment: 입력 영상의 크기에 대한 전처리 방법
  + Padding: 2D 컨볼루션 연산을 위한 전처리 방법

1. **소요되는 주요 연구시설 및 기기**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 주요시설 및 기기 | 규격 | 유무 | 설치기관 | 없는 경우의 대책 |
| Linux Server  (4 GPUs) | 600 x 800 x 500  mm | 有 | 인천대학교 전자공학과  SoC 연구실 |  |
| S2C singleE Virtex7  TAI Logic Module | 200 x 150 x 30  mm | 有 | 인천대학교 전자공학과  SoC 연구실 |  |
| ZedBoard | 150 x 100 mm | 有 | 인천대학교 전자공학과  SoC 연구실 |  |

1. **연구일정**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 월  항목 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 자료수집 |  |  |  |  |  |  |  |  |  |  |  |  |
| 이론 및 기존 연구 분석 |  |  |  |  |  |  |  |  |  |  |  |  |
| 구조 설계 및 시뮬레이션 |  |  |  |  |  |  |  |  |  |  |  |  |
| 제약 조건 최적화 |  |  |  |  |  |  |  |  |  |  |  |  |
| 결론 도출 |  |  |  |  |  |  |  |  |  |  |  |  |
| 논문 작성 및 수정 |  |  |  |  |  |  |  |  |  |  |  |  |

1. **기대되는 성과 및 활용 방안**
   * 심층 CNN이 구현된 뉴로모픽 시스템의 성능 및 에너지 효율성 증대
     + 저전력 인공지능 기반 반도체 회로 설계 기술 확보
     + 높은 수준의 에너지 효율성 확보로 모바일 시스템에 적용 가능한 칩 개발
   * 다양한 분야에 적용 가능한 Flexible 설계 방법 확보
     + FPGA-Based 혹은 ASIC-based 에 적용 가능한 뉴로모픽 시스템 설계 기술
2. **참고문헌**

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

[2] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

[3] Jeff Parkhurst, John Darringer, and Bill Grundmann. From single core to multi-core: preparing for a new exponential. In Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design, pages 67–72. ACM, 2006.

[4] Doug Burger, James R Goodman, and Alain K¨agi. Memory bandwidth limitations of future micro-processors, volume 24. ACM, 1996.

[5] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. cience, 345(6197):668–673, 2014.

[6] Jaeyong Chung, Taehwan Shin, and Yongshin Kang. Insight: A neuromorphic computing system for evaluation of large neural networks. arXiv preprint arXiv:1508.01008, 2015.

[7] Keshab K Parhi. VLSI digital signal processing systems: design and implementation. John Wiley & Sons, 2007.

[8] Yu-Chi Tsao and Ken Choi. Area-efficient vlsi implementation for parallel linear-phase fir digital filters of odd length based on fast fir algorithm. IEEE Transactions on Circuits and Systems II: Express Briefs, 59(6):371–375, 2012.

[9] Praveen K Murthy and Edward A Lee. Multidimensional synchronous dataflow. IEEE Transactions on Signal Processing, 50(8):2064–079, 2002.

[10] Jaeyong Chung and Taehwan Shin. Simplifying deep neural networks for neuromorphic architectures. In Proceedings of the 53rd nnual Design Automation Conference, 2016.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.