IDEC Chip Design Contest

본 연구는 IDEC에서 MPW를 지원받아 수행하였습니다.

CTCAM 기반 SFNN 이미지 분류 알고리즘의 Verilog 구현

김동휘, 김수민, 임현기, 서영석, 최연우, 허정민*, 홍상훈
경희대학교 전자공학과
경희대학교 반도체융합학과*

본 연구에서는 CTCAM(Computational Ternary Content-Addressable Memory)을 기반으로 SLDP(Spike Location Dependent Plasticity)학습이 가능한 SFNN(Search Frequency Neural Network) 이미지 분류 알고리즘 제시한다. 이는 고속 이미지 분류를 위한 DPIM(Digital Processing In Memory) 구조로, Verilog HDL을 통해 구현 및 검증하였다. 제안한 구조는 엣지 필터링과 커브 필터링 기법을 통해 입력 이미지로부터 주요 시각적 특징을 추출하며, 이 과정에서 CTCAM을 통해 고빈도 패턴을 학습하고 이를 필터링에 사용함으로써 효율적인 특징 추출을 가능하게 하였다. 또한, 출력층에서 위치별 패턴의 중요도를 학습하고 추론 과정에서 이를 반영해 확률 기반의 신뢰도를 제공한다.

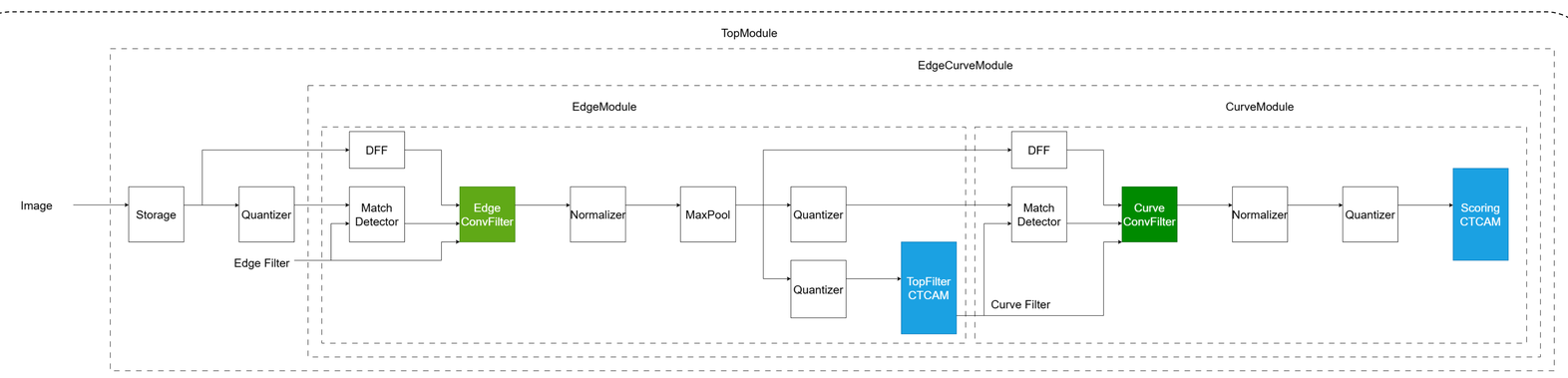


Fig. 1. Overall Block Diagram

본 연구에서는 피드포워드 방식의 학습 알고리즘을 사용하는 디지털 메모리 내 처리(DPIM: Digital Processing In Memory) 구조를 제안한다. 이 방식에서는 이전 계산 단계에 영향을 주는 연산 단계가 존재하지 않는다. 우리는 이미지 분류에 사용하기 위한 새로운 학습 메커니즘으로 STDP(Spike-Timing Dependent Plasticity)와 유사한 스파이크 위치 의존 가소성(SLDP, Spike Location Dependent Plasticity) 방식을 사용한다. 제안된 방법에서는, TCAM(Ternary Content Addressable Memory)을 활용하며, 이는 검색 패턴이 입력되면 저장된 주소를 출력하는 방식의 메모리이다. TCAM에서 자주 검색되는 고빈도 패턴들이 더 높은 주소로 재배치되도록 구성하면, 입력 패턴을 우선순위에 따라 정렬할 수 있게 되어 학습이 이뤄진다.

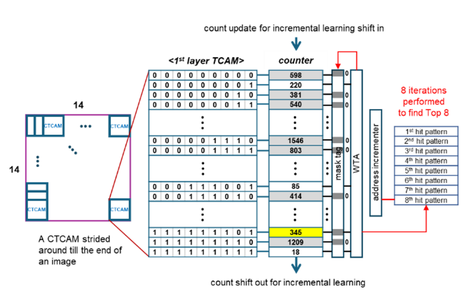


Fig. 2. 1st CTCAM Layer

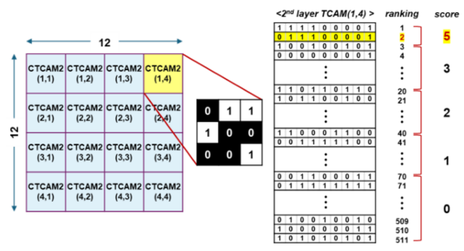


Fig. 3. 2nd CTCAM Layer

첫 번째 학습이 이뤄지는 층에서 CTCAM은 입력 이미지에 대해 3x3 패턴을 추출해 패턴 빈도를 학습한다. 이 과정은 Fig.2에 나타나 있다. 두 번째 학습이 이뤄지는 층에서는 입력 패턴에 해당하는 블록 공간이 최종 결과에 미치는 영향을 강화하거나 약화시키는 방식으로 학습이 진행된다. 전체 이미지 공간은 이러한 블록 공간들로 구성되며, 모든 블록의 영향력을 합산한 값이 해당 입력 이미지의 표현값이 된다. 따라서 특정 클래스를 대표하는 이미지들이 SLDP에 반복적으로 입력될 경우, 유사한 특징들이 강화되며, 최종 합산값이 증가하게 된다. 이 과정은 Fig.3에 나타나 있다. 이전 시점으로 되돌아가며 업데이트를 추적할 필요가 없기 때문에, 이 학습 방식은 완전히 피드 포워드 구조로 동작한다.

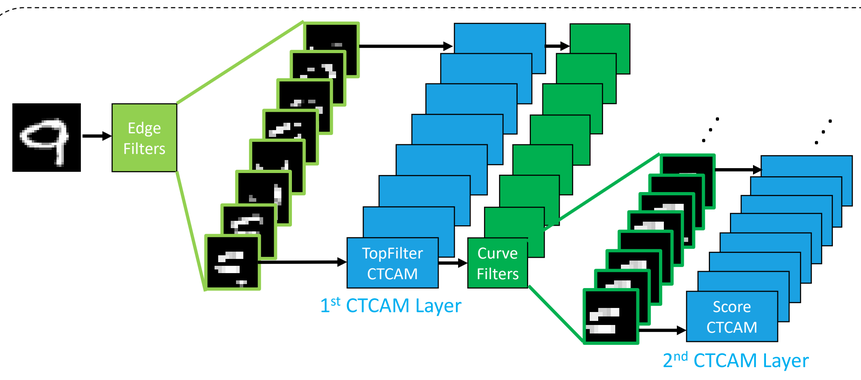


Fig. 4. Image Visualization Across Layers

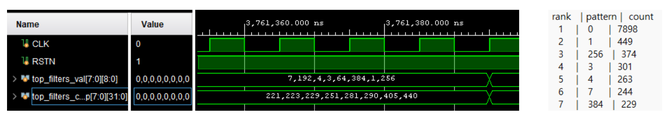
실제 학습 과정에서 단계별 내부 이미지를 Fig.4에 나타내었다. 전체 하드웨어 구조는 이미지 전처리부터 최종 분류까지의 전 과정을 처리할 수 있도록 TopModule을 중심으로 구성되어 있다. 입력 이미지가 순차적으로 엣지 필터링, 양자화, 맥스 풀링, 커브 필터링을 거쳐 CTCAM 기반의 스코어링으로 연결되는 파이프라인 구조를 갖추고 있으며, 각 단계는 독립적인 서브 모듈로 설계되어 병렬 처리를 지원한다. 모든 데이터 처리는 동기식 클럭에 기반하며, 실시간 학습과 추론이 가능하도록 구성되어 있다. 상위 FSM(Finite State Machine) 모듈은 Top Module, EdgeCurve Module, Edge Module, Curve Module로 구분된다. Top Module은 전체 제어를 담당하며, 각각의 상태에 대한 신호를 통해 전이되며 각 처리 흐름을 제어한다. 모듈별 연산은 Fig.1에서 확인할 수 있다.

추론 단계에서는 학습된 Top-8 필터를 기반으로 커브 필터링이 수행된다. Scoring CTCAM 모듈은 학습 시 CTCAM의 패턴의 카운트에 따라 정렬을 진행하며, 추론 시 랭크 정보를 기반으로 입력 패턴에 점수를 부여하며, 그 점수를 출력한다. 이렇게 출력된 각 score은 Top Module에서 누적되어 최종적으로 모두 합해진다. 이 누적된 점수는 개별 class에 대한 최종 예측 score이다. 모든 class에서 최종 예측 score가 산출되면 가장 큰 score을 가진 class를 최종 예측으로 결정한다.

제안된 구조는 학습 단계에 파이프라인 방식을 도입하여, 이미지를 입력 받는 단계를 이후 단계들과 병렬적으로 연산 가능하도록 설계하였다. 이를 통해 입력 데이터 수신과 동시에 Quantization 및 Edge · Curve Filtering 연산이 중첩 실행되어 전체 학습 시간을 효과적으로 단축할 수 있었다. 파이프라인 구조에서는 State 1의 794 사이클이 전체에서 단 1회만 소요되며, 이후 각 학습 샘플에 대해 State 2(794 사이클)와 State 3(1,935 사이클)만 반복되므로, 6,000장의 학습 이미지 기준 총 연산량은 16,374,794 사이클로 계산된다.

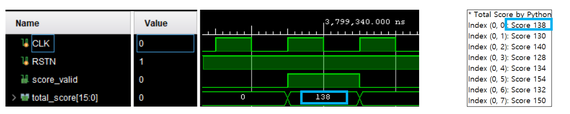
Xilinx Vivado를 활용하여 제안된 하드웨어 구조에서 학습 및 추론 시뮬레이션을 수행해 Python 소프트웨어 모델과의 비교를 통해 유효성을 검증하였다.

Fig.5, 6에 시뮬레이션의 일부 결과가 나타나 있다.



(a) Verilog (b) Python

Fig. 5. TopFilterCTCAM Sorting Result



(a) Verilog (b) Python

Fig. 6. ScoreCTCAM Score Output

본 연구에서는 CTCAM(Computational Ternary Content-Addressable Memory) 기반 구조를 활용하여 CNN과 유사한 단계적 이미지 처리 방식으로 적용 가능한 고속 DPIM(Digital Processing-In-Memory) 하드웨어를 설계하였다. 실제 MNIST 데이터셋을 대상으로 제안된 하드웨어 구조의 성능을 검증한 결과, 94.4%의 분류 정확도를 달성하였다. 연산 효율성 측면에서 두드러진 향상이 관찰되었는데, 기존의 SE-SFNN(Stride Edge-detection Search Frequency Neural Network) 구조가 학습을 완료하는 데 총 423,576,800 사이클이 요구된 반면, 제안된 CTCAM 기반 DPIM 구조는 단 16,374,794 사이클만으로 동일 작업을 수행하였다. 이는 약 25.9배에 달하는 사이클 수 감소로, 학습 시간 측면에서 현저한 최적화가 이루어졌음을 시사한다. 이와 같은 결과는 CTCAM이 단순한 검색 메모리를 넘어 학습과 추론을 모두 효율적으로 수행할 수 있는 뉴로모픽 하드웨어 구현에 효과적으로 활용될 수 있음을 보여주며, 제안된 구조가 저전력·고속 연산이 요구되는 엣지 컴퓨팅 환경이나 실시간 이미지 분류 시스템에 적용될 수 있는 가능성을 제시한다.