논문 2021-2-17 http://dx.doi.org/10.29056/jsav.2021.12.17

뉴로모픽 구조 기반 FPGA 임베디드 보드에서 이미지 분류 성능 향상을 위한 특징 표현 방법 연구

정재혁*, 정진만**, 윤영선*†

Feature Representation Method to Improve Image Classification Performance in FPGA Embedded Boards Based on Neuromorphic Architecture

Jae-Hyeok Jeong*, Jinman Jung**, Young-Sun Yun*†

요 약

뉴로모픽 아키텍처는 저에너지로 인공지능 기술을 지원하는 차세대 컴퓨팅으로 주목받고 있다. 그러나 뉴로모픽 아키텍처 기반의 FPGA 임베디드 보드는 크기나 전력 등으로 인하여 가용 자원이 제한된다. 본 논문에서는 제한된 자원을 효율적으로 사용하기 위해 특징점의 고려 없이 크기를 재조정하는 보간법과 에너지 기반으로 특징점을 최대한 보존하는 DCT(Discrete Cosine Transform) 기법을 통한 특징 표현 방법을 비교 및 평가한다. 크기가 조정된 이미지는 일반적인 PC 환경에서와 FPGA 임베디드 보드의 Nengo 프레임워크에서 컨벌루션 신경망을 통해 정확도를 비교 분석했다. 실험 결과 PC의 컨벌루션 신경망과 FPGA Nengo 환경 모두에서 DCT 기반 분류 성능이 일반 보간법보다 약 1.9% 높은 성능을 보였다. 실험 결과를 바탕으로 뉴로모픽 구조 기반 FPGA 보드의 제한된 자원 환경에서 기존에 사용되던 보간법 대신 DCT 방식을 이용한다면 분류에 사용되는 뉴런의 표현에 많은 자원을 할당하여 인식률을 높일 수 있을 것으로 기대한다.

Abstract

Neuromorphic architecture is drawing attention as a next-generation computing that supports artificial intelligence technology with low energy. However, FPGA embedded boards based on Neuromorphic architecturehave limited resources due to size and power. In this paper, we compared and evaluated the image reduction method using the interpolation method that rescales the size without considering the feature points and the DCT (Discrete Cosine Transform) method that preserves the feature points as much as possible based on energy. The scaled images were compared and analyzed for accuracy through CNN (Convolutional Neural Networks) in a PC environment and in the Nengo framework of an FPGA embedded board.. As a result of the experiment, DCT based classification showed about 1.9% higher performance than that of interpolation representation in both CNN and FPGA nengo environments. Based on the experimental results, when the DCT method is used in a limited resource environment such as an embedded board, a lot of resources are allocated to the expression of neurons used for classification, and the recognition rate is expected to increase.

한글키워드 : 뉴로모픽 아키텍처, 이산 코사인 변환, 이미지 재조정, 임베디드 뉴로모픽 보드

keywords : Neuromorphic architecture, Discrete Cosine Transform, Image Reduction, Embedded Neuromorphic Boards

* 한남대학교 정보통신공학과
** 인하대학교 컴퓨터공학과
† 교신저자: 윤영선(email: ysyun@hnu.kr)

접수일자: 2021.11.29. 심사완료: 2021.12.06.
게재확정: 2021.12.20.

1. 서 론

최근 인공지능 기술에 기반한 IoT 기술이 관심을 받고 있으며, 그 활용이 늘어나고 있다. 각광을 받고 있는 딥러닝 기반 인공지능 기술은 지금까지 일반 컴퓨터가 취약했던 패턴인식, 자율주행, 게임, 추론 등 다양한 방면에서 널리 활용되고 있으나, 높은 성능을 보이기 위해서는 고성능 컴퓨팅 환경에서 학습에 많은 시간과 전력을 소비하고 있다. 실제로 2016년 이세돌과 바둑을 둔 알파고의 경우 대국을 두던 당시 이세돌이 소비한 약 20W에 비하여 약 1MW가량의 전력을 소모한 것으로 알려져 있으며, 이는 컴퓨터가 사람보다 약 50,000배 이상 또는 일반 가정집 100가구가 하루 소비하는 전력을 사용한 것으로 알려져 있다[1]. 이런 이유로 컴퓨터와 비교하여 매우 효율적인 에너지 구조를 가지는 인간의 뇌를 모방하여 인공지능의 성능을 유지하면서 에너지 소모를 줄이려는 뉴로모픽 구조 연구 등이 활발히 진행되고 있다[2].

본 논문에서는 소규모 IoT 환경에서 저전력 기반으로 일반 인공지능의 성능을 보이고자 시도하는 뉴로모픽 기반의 내장형 보드 환경에서 대표적인 이미지 분류 연구의 성능 향상을 기술한다. 연구에 사용한 플랫폼은 뉴로모픽 구조를 FPGA로 구현한 보드이며, 현재 프로토타입으로 다양한 종류의 임베디드 보드가 출시되어 연구중이다. 대표적인 임베디드 보드로는 Xilinx사의 PYNQ, Intel 사의 DE1-SoC, Loihi칩 등이 있으며, 본 논문에서는 PYNQ와 DE1-SoC에서 뉴로모픽 개발 플랫폼인 nengo 기반으로 스파이크 신경망(Spiking Neural Network, SNN)[3]을 적용하여 연구를 진행하였다. 현재 출시된 뉴로모픽 구조 기반의 FPGA 임베디드 보드는 크기와 전력 소비 등을 고려할 때 설계적 제약이 발생하며, 사용할 수 있는 자원이 제한된다. 따라서 이미지 분류 연산의 경우 이미지의 해상도가 높으면 모델링 할 수 있는 뉴런(자원)의 수가 충분하지 않아 정확도가 낮아지는 결과를 초래한다[4]. 따라서 기존의 뉴로모픽 구조 기반 FPGA 보드 환경의 이미지 분류 연구에서는, 성능 향상을 위하여 입력 뉴런의 수를 줄임으로써, 연산 및 모델링에 사용되는 분류 뉴런의 수를 더 많이 확보하여 정확도를 높이는 방안을 고려하였다.

입력 뉴런을 줄이기 위한 일반적인 방법은 회색 이미지나, 흑백 이미지로 변환하는 것처럼 색상을 표현하는 채널 수를 줄이거나, 이미지의 해상도를 줄이는 방법이 사용된다. 회색 이미지나 흑백 이미지로 변환하는 방법은 수집된 이미지가 회색이나 흑백인 경우 효과가 없기 때문에, 본 연구에서는 이미지의 해상도를 줄이는 방법에 관한 연구를 진행하였다. 이미지의 해상도를 낮추는 방법은 일반적으로 이미지 보간법이 사용되며 대표적인 방식으로는 Nearest neighbor, B-linear, Area, Bicubic, Lanczos 기법 등이 있다. 이러한 보간법들은 이미지의 특징점을 고려하지 않고 단순히 해상도를 조절하여 임의의 화소 값을 삽입하거나 제거하는 방식으로 동작하기 때문에 aliasing, 열화 현상 등이 발생하여 이미지의 왜곡 등으로 정확도가 떨어질 수 있기 때문에, 본 연구에서는 이미지 압축에 널리 사용되는 이산 코사인 변환(discrete cosine transform, DCT)을 이용하여, 압축된 에너지를 입력 특징으로 최대한 이미지의 주요 특징점을 보존하고자 하였다[5]. 에너지 압축은 이미지의 전체적인 특징을 표현하는 낮은 주파수 특징을 보존하고, 이미지의 개별 특징이나 변화에 민감한 높은 주파수 특징을 분리하여 이미지의 특징 정보를 효율적으로 표현할 수 있도록 DCT의 결과에서 저주파 부분을 추출하여 이용하는 방법이다. 이산코사인 기반의 특징 표현 방법은 에너지를 압축하여 에너지가 응축된 부분을 보존하기 때문에 이

미지의 주요 특징점을 보존할 수 있다는 장점이 있지만, 이미지의 해상도가 높을수록 계산량이 많아지는 단점도 존재한다. 그러나 최근 소형 전용 디지털 신호 처리 장치나 전용 그래픽 처리 장치 등이 활발히 개발되고 있으므로, DCT 기반의 분류 시스템의 성능이 확인된다면 추후 연산량 감소 등의 연구를 통하여 실제 환경에 적용할 수 있을 것으로 판단한다. DCT 기반의 연산량을 줄이는 연구로는 양자화기[6], SAD 정보를 이용한 DCT 계산방식[7] 등이 있다. 제안된 방법의 유효성을 확인하기 위하여 본 논문에서는 제한된 FPGA 보드 자원 크기에 맞게 다양한 보간법을 이용하여 축소된 이미지와 DCT 방법에 의하여 표현된 이미지를 이용하여 일반적인 심층 신경망과 뉴로모픽 개발 환경에서 정확도를 비교, 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 뉴로모픽 신경망을 개발하는 Nengo 프레임워크에 관하여 소개한다. 3장에서는 이미지 표현 방법인 보간법과 DCT에 관하여 기술한다. 4장에서는 보간법과 DCT 기법을 통해 뉴로모픽 구조 기반 FPGA 임베디드 보드에 적용 가능한 이미지 표현 방법과 정확도를 도출하는 방법을 제안한다. 5장에서는 원본 이미지, 보간법 그리고 DCT의 실험 결과를 도출했다. 6장에서 결론을 맺는다.

2. 뉴로모픽 신경망 개발 프레임워크

본 논문은 널리 알려진 합성곱 신경망(convolutional neural network, CNN)을 적용한 심층 신경망과 뉴로모픽 아키텍처인 SNN에 특화된 제한된 자원 환경에서의 이미지 축소 및 에너지 압축에 의한 특징 표현 방법을 비교한다. 이번 장에서는 합성곱 신경망의 경우 널리 알려진 기술이기 때문에 관련 소개는 생략하고 뉴로모픽 아키텍처에서 신경망을 구현하는 Nengo 프레임워크에 대하여 소개한다.

뉴로모픽 아키텍처는 인간의 뇌를 모방하여 동작하며 코어가 뉴런의 역할을 담당하고 메모리 칩이 뉴런을 이어주는 시냅스 역할을 담당하는 구조이다. 뉴로모픽 아키텍처를 대표하는 SNN 모델은 두뇌에서 실제로 정보가 전달되고 가공되는 과정을 모방하여 뉴런과 시냅스로 인공지능을 구현한다. 두뇌는 신호가 전달된 뉴런만을 활성화시키기 때문에 모든 신경망에서 연산을 수행하는 기존의 인공신경망에 비하여 저전력의 수행이 가능하다. 그림 1과 같이 SNN 모델에서는 각 뉴런은 입력(입력 뉴런, Input)으로부터 이산적인 스파이크 신호를 만들고, 시냅스를 통하여 전달된 신호를 누적하여 임계치를 넘어서면 (흥분(excitatory) 뉴런, Exc) 다른 뉴런으로 스파이크 신호를 전달하여 측면 억제를 유도 (억제(inhibitory) 뉴런, Inh)하는 식으로 동작한다. 자극에 따라 세포막의 전위차를 모델링하는 뉴런 모델은 LIF(Leaky Integrate and Fire), Izhikevich, Hodgkin-Huxley 등이 사용되며, Nengo 프레임워크를 통해서 뉴로모픽 기반의 알고리즘을 쉽게 개발할 수 있다.

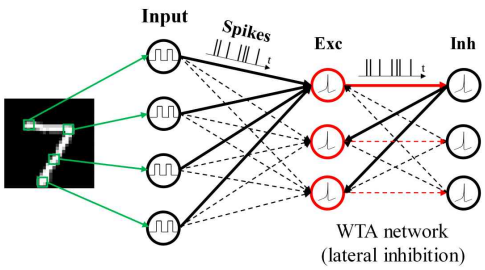


그림 1. 스파이킹 신경망의 구조[8].
Fig. 1. The architecture of a spiking neural network[8].

Nengo[9]는 사람의 뇌를 모방하기 위하여 생물학적으로 대규모 인지 신경 모델을 구축할 수 있는 신경 공학 프레임워크[10]를 사용하여 신경망을 생성하기 위한 오픈 소스 기반의 신경 시뮬

레이터이다. Nengo는 입력을 받아들이는 입력층, 연산을 수행하는 은닉층, 결과를 출력하는 출력층으로 구성되며, 뉴로모픽 하드웨어 환경에서뿐만 아니라 뉴로모픽 하드웨어가 없는 환경에서도 SNN 모델을 사용할 수 있도록 시뮬레이션 환경을 제공한다. 또한, Nengo 프레임워크는 기존의 신경망 프레임워크와 연동하여 사용될 수 있도록 Tensorflow 기반의 NengoDL 라이브러리와, FPGA 보드 환경에서 동작 가능한 NengoFPGA 라이브러리를 제공하고 있다.

3. 입력 특징 표현

입력 이미지를 Nengo의 입력층에 전달하기 위하여 FPGA의 자원에 적합한 이미지로 축소하거나, 입력 특징을 변환하여야 한다. 본 장에서는 대표적인 이미지 축소 방법인 보간법과 이산 코사인 변환을 통한 입력 특징 표현에 대하여 설명한다.

보간법은 특징(화소)값들을 이용하여, 임의의 값들을 생성하는 방법으로 주변 화소의 값을 기반으로 최상의 근사치를 얻어내는 과정이다. 반면 DCT 기법은 입력 신호를 주파수 성분별 크기로 변환하는 방식으로 신호 처리에서 널리 사용되는 푸리에 변환과 비슷한 원리로 구현된다.

3.1 보간법

최근접 이웃 보간법(nearest neighbor)[11]은 이미지를 확대하거나 축소할 경우 가장 가깝게 이웃한 원시 화소의 값을 할당하는 방법이다. 이 보간법은 기존 이미지의 화소를 반복시키거나 이웃한 화소를 선택하기 때문에 가장 빠른 방법이지만 이미지 계단 현상이나 특징점 소실 등의 단점이 존재한다.

1차원 선형 보간법은 두 값이 주어졌을 때 그 사이에 있는 값을 추정하기 위하여 직선거리에 따라 선형적으로 계산하는 방법이다. 쌍 선형 보간법(Blinear)[12]은 1차원 선형 보간법을 확장하여 2차원 데이터에 적용하는 방법이다. 쌍 선형 보간법은 그림 2와 같이 네 점의 데이터 값 A, B, C, D에서 임의의 점 P에서의 보간 값은 거리 값 h_1, h_2, w_1, w_2에 따라 식 (1)과 같이 계산하여 사용한다.

$$\alpha = \frac{h_1}{h_1 + h_2}, \quad \beta = \frac{h_2}{h_1 + h_2}, \quad p = \frac{w_1}{w_1 + w_2}, \quad q = \frac{w_2}{w_1 + w_2}$$

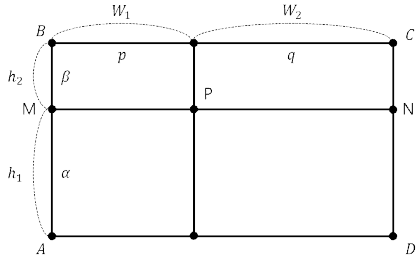


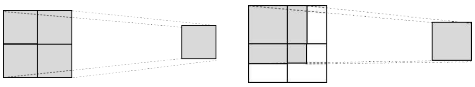
그림 2. 쌍선형 보간법
Fig. 2. Bilinear interpolation

$$P = q(\beta A + \alpha B) + p(\beta D + \alpha C) \tag{1}$$

계산 원리는 A, B를 보간하여 M의 값을 얻고 C와 D를 보간하여 N의 값을 얻은 후 M과 N을 보간하여 P의 값을 도출한다. 쌍 선형 보간법을 이용한 이미지 축소는 그림 3과 같이 축소하고자 하는 해상도가 정수배 또는 실수배인지에 따라 식(1)을 통해 화소 값을 계산한다. 정수배 축소의 경우에는 해당되는 영역을 균등하게 반영하여 평균을 적용하며, 실수배의 경우에는 해당되는 영역에 해당하는 화소값을 비율에 따라 반영하여 평균 값으로 산출한다.

Area 영역 보간법은 쌍 선형 보간법을 이용하여 영역의 면적을 이용하는 방법이다. 그림 3(a)와 같이 원본 이미지를 1/4로 축소하는 경우 정수배에 해당하므로 화소 4개에 대한 평균을 구하여 계산한다. 하지만 그림 3(b)와 같이 실수 배로 축소시킬 경우, 각 화소의 가중치 1.0, 0.5, 0.5,

0.5*0.5와 총합 2.25를 반영하여 가중 평균값을 통해 화소 값을 산출한다.



(a) Reduced with on integer ratio (b) Reduced with on float ratio

그림 3. 이미지 축소 예시
Fig. 3. Examples of image downsizing

쌍 삼차원 보간법(Bicubic)[13]은 그림 4와 같이 알려진 16개의 점 데이터 값을 통해 수직으로 식 (2)를 통해 4개의 x_1, x_2, x_3, x_4좌표를 구한 후 수평으로 한번 더 보간하여 새로운 좌표 p_1를 생성한다. 여기에서 α의 값은 일반적으로 -0.5, -0.75, -1을 사용한다.

$$f(x)=\begin{cases}(\alpha+2)|x|^3-(\alpha+3)|x|^2+1, & 0\le|x|<1\\ \alpha|x|^3-5\alpha|x|^2+8\alpha|x|-4\alpha, & 1\le|x|<2\\ 0, & 2\le|x|\end{cases} \quad (2)$$

쌍 삼차원 보간법은 쌍 선형 보간법보다 많은 점 좌표를 요구하므로 화질 면에서 우수하지만 더 많은 계산량을 요구한다. 확대의 경우 그림 4와 같이 계산하고, 축소의 경우 그림 3과 같이 정수배와 실수배에 관하여 식(2)를 통해 화소 값을 계산한다.

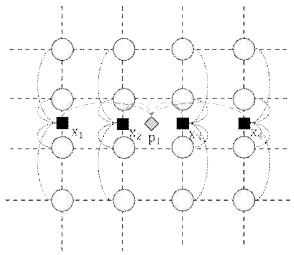


그림 4. 쌍 삼차원 보간법
Fig. 4. Bicubic interpolation

Lancozs[14] 보간법은 Reconstruction Kernel 방정식인 L(x)를 가중치의 함수로써 사용한다. L(x)의 수식은 식 (3)과 같다. 여기서 α의 값은 일반적으로 2 또는 3으로 사용한다.

$$L(x)=\begin{cases}\mathrm{sinc}(x)\times\mathrm{sinc}(\frac{x}{a}), & -\alpha<x<\alpha\\ 0, & otherwise\end{cases}$$

$$L(x)=\begin{cases}1, & x=0\\ \frac{\alpha\sin(\pi x)\sin(\frac{\pi x}{\alpha})}{\pi^2x^2}, & 0<|x|<\alpha\\ 0, & otherwise\end{cases} \quad (3)$$

이 식을 사용하여 1차원의 신호 값은 식 (4)과 같으며 2차원의 신호로 표현하면 식 (5)과 같이 표현된다.

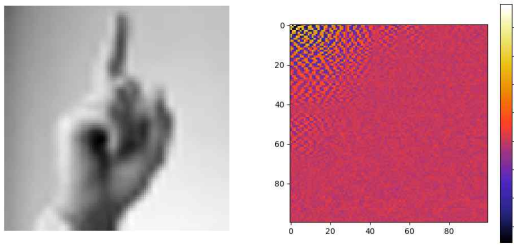
$$S(x)=\sum_{i=\lfloor x\rfloor-\alpha+1}^{\lfloor x\rfloor+\alpha}S_iL(x-i) \quad (4)$$

$$S(x,y)=\sum_{i=\lfloor x\rfloor-\alpha+1}^{\lfloor x\rfloor+\alpha}\sum_{j=\lfloor y\rfloor-\alpha+1}^{\lfloor y\rfloor+\alpha}S_{ij}L(x-i)L(y-j) \quad (5)$$

이러한 Lancozs 보간법은 선명한 화질을 제공하지만 작은 이미지에 대해서는 노이즈까지 부각시키기 때문에 화질 열화가 더 두드러지는 단점이 있다.

3.2 이산 코사인 변환

이산 코사인 변환 (DCT)[15]는 코사인 기저함수를 사용하는 변환 방법이다. 기저함수로 지수함수를 사용하는 푸리에 변환에 근거하여 주기가 다양한 주기의 코사인 함수로 공간 정보를 주파수 영역으로 변환하는 방법이다. 이미지 및 영상에서 인접한 화소 간에는 비슷한 색상인 경우가 많기 때문에 그림 5와 같이 원본 이미지를 DCT 하였을 경우 화소의 값의 변화가 적은 경우, 코



(a) Original Image (b) Log DCT coefficient

그림 5. 에너지 집중 현상 예시
Fig. 5. Example of energy concentration

사인 계수 값이 낮은 주파수로 몰리게 되며 이를 에너지 집중 현상이라 한다. 반대로 화소의 변화가 큰 경우에는 높은 주파수에 위치한다. 에너지 집중 현상을 통해 이미지를 표현하는 경우, 낮은 주파수 성분에는 원래 이미지의 전역 성분을 많이 포함하지만 높은 주파수 영역에는 지역적인 성분이 표현되어, 높은 주파수 영역을 제거하여도 이미지 특성의 차이가 크게 관찰되지 않는다. 따라서 DCT에 의한 에너지 압축 방법을 이용한다면 이미지의 전체크기에 대하여 에너지가 집중된 영역만을 이용할 수 있다. 그림 5(b)는 이미지를 DCT 한 후 값 변화를 쉽게 관찰하기 위하여 DCT 계수에 로그 함수를 적용한 후 에너지 집중도를 보이고 있다. 그림 5(b)에서 보는 바와 같이 좌표 (0,0)에 에너지가 집중된 현상을 볼 수 있으며 최초의 좌표 (0,0)에는 전체 이미지의 평균 크기인 DC 성분이 위치한다.

4. FPGA 보드에 적용 가능한 이미지 표현기법

광학 기술의 발전에 따라 이미지의 해상도는 증가되고 있으나 뉴로모픽 아키텍처 기반의 FPGA 임베디드 보드에 적용 가능한 자원은 메모리나 연산 처리기 등에서 매우 제한적이다. 해상도 높은 이미지를 축소 없이 입력으로 제공하는 경우, 보드의 메모리 등 제약으로 인하여 연산에 필요한 메모리가 부족해져 충분한 수의 뉴런을 모델링 하지 못하여 정확도가 떨어진다. 따라서 본 논문에서는 제한된 FPGA 자원에서 해상도를 낮추어 입력 특징의 수를 줄이고 연산에 필요한 뉴런을 확보하기 위하여 보간법과 DCT 기법을 활용한 이미지 특징 표현 및 전처리 방법을 제안한다.

연구에 사용한 데이터셋은 28x28이미지로 구성된 Digits MNIST, Fashion MNIST, Sign Language MNIST 3가지를 사용하였다. MNIST 데이터셋은 심층 신경망의 성능을 평가하기 위하여 일반적으로 사용되는 데이터로서, 단순한 필기 숫자에서부터 패션, 실생활에 적용 가능한 수화이미지를 포함하고 있으며, 점차 이미지의 복잡도가 높아지는 경향을 보인다. 본 연구에서는 각 이미지의 복잡도에 따라 이미지 특징 표현 방법의 성능 변화를 파악하기 위하여 MNIST 데이터셋을 이용하였다. 임베디드 보드에서 사용한 입력 특징 벡터로는 14x14로 설정한다. 이미지 자체를 이용하는 경우 28x28 이미지를 14x14로 축소하였으며, DCT의 경우에는 14x14 계수를 이용한다. 입력 특징 표현을 위하여 5가지의 보간법(Resize)과 DCT를 이용하여 이미지 축소와 특징 벡터를 계산하였으며, 그림 6과 같이 CNN을 통과하여 각 이미지를 클래스별로 분류한다.

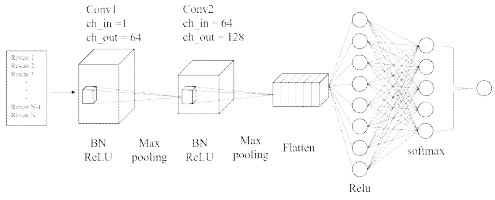
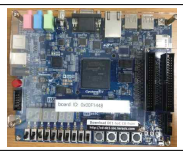
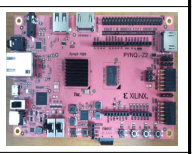


그림 6. CNN 모델 구조
Fig. 6. CNN model architecture

실험에 사용된 합성곱 신경망 구조는 활성화 함수로 ReLU를 사용하는 2개의 합성곱 층에 대하여 각각 batch normalization[16] 과정과 max pooling의 과정을 수행한다. 합성곱 결과는 평탄화(flatten) 과정을 거쳐 최종적으로 softmax 함수에 의해 해당되는 클래스를 확률값으로 표현한다.

자원이 제한된 FPGA 환경에서는 그림 7과 같이 모델을 구축하여 결과를 측정한다. 실험에 사용된 뉴로모픽 아키텍처 기반 FPGA 임베디드 보드는 표 1과 같이 DE1-SoC 보드는 800MHz를 지원하는 듀얼코어 기반 ARM Cortex-A9 칩을 사용하여 메모리는 1GB로 처리 속도는 PYNQ 보드에 비하여 빠르지만 모델링 할 수 있는 뉴런의 수가 16K로 한정되어 있기 때문에 정확도가 낮다. 반면에 PYNQ 보드의 경우 650MHz를 지원하는 듀얼코어 기반 ARM Cortex-A9 칩을 사용하고 메모리는 512MB이지만 모델링 할 수 있는 뉴런의 수가 32K이기 때문에 연산 속도 측면에서는 DE1-SoC 보드보다 조금 느리지만 동일한 데이터셋의 경우 조금 더 높은 정확도를 도출한다.

표 1. FPGA 보드 성능
Table 1. FPGA boards specification

	DE1-SoC	PYNQ
Board		
Processor	800MHz dual-core ARM Cortex-A9	650MHz dual-core ARM Cortex-A9
Memory	1GB	512MB
Max No. of Neuron	16K	32K

NengoFPGA 네트워크 모델은 그림 7과 같이 입력층, 은닉층, 출력층으로 구성된다. 입력층은 입력 데이터에 임의의 수를 곱하여 내장 벡터로 표현하는 단계로, 14x14 이미지 또는 DCT 계수로 표현된 총 196개의 뉴런이 사용된다. 은닉층은 입력층의 데이터를 이용하여 실제로 연산을 수행하는 부분이다. FPGA 보드에서는 자원의 제약으로 하나의 은닉층만을 사용할 수 있으며 은닉 뉴런의 수는 가용 가능한 총 뉴런의 수를 입력층에 할당된 뉴런의 수로 나눠 결정한다. 따라서 입력층에 전달되는 입력 특징 벡터의 수가 커지면 사용 가능한 뉴런의 수는 줄어들게 된다. DE1-SoC의 경우 사용할 수 있는 총 가용 뉴런의 수가 16K이며 입력층에 사용된 뉴런이 196개이기 때문에 총 81개의 은닉 뉴런을 사용할 수 있으며, PYNQ의 경우 162개의 은닉 뉴런을 사용할 수 있다. 마지막으로 출력 레이어는 은닉층에서 계산된 가중치를 통합하여 정답을 추론한다.

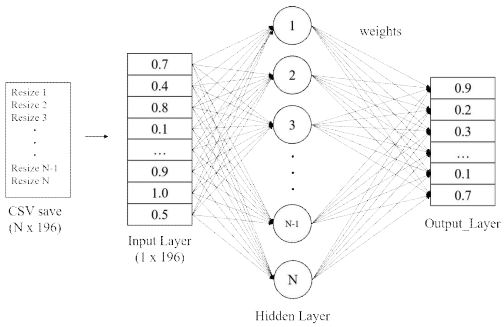


그림 7. Nengo FPGA 네트워크 모델 구조
Fig. 7. Nengo FPGA Network model structure

5. 실험 결과

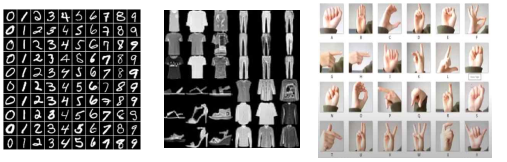
5.1 실험 환경

실험은 Nvidia RTX 2070이 장착된 Windows 환경에서 진행하였고, 표 1과 같이 자원이 제한된 FPGA 환경인 DE1-Soc와, PYNQ 보드에서 실험을 진행하였다. 실험에 사용한 데이터셋의

표 2. 데이터셋에 대한 학습 및 테스트 조건
Table 2. Train and test conditions for each data set

MNIST Dataset	Digits	Fashion	Sign Language
Train	60,000	60,000	27,455
Test	10,000	10,000	7,172

이미지는 모두 28x28 이미지로 구성되었으며, 데이터의 예는 그림 8에 표시하였다.



(a) Digits MNIST (b) Fashion MNIST (c) Sign Language MNIST

그림 8. 데이터셋 예시
Fig. 8. Examples of data sets

Digits MNIST는 손으로 쓰여진 0~9까지의 데이터셋을 의미하며 표 2와 같이 6만장의 train 데이터, 1만장의 test 데이터셋으로 구성되어있다. Fashion MNIST는 T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankel boot로 총 10가지의 클래스로 구성되었으며 Digits MNIST와 동일하게 6만장의 train 데이터, 1만장의 test 데이터셋으로 구성되어있다. 마지막으로 Sign Language MNIST는 A~Z까지 총 26가지의 알파벳 수화를 표현한 것이며 train 데이터 27,455장, test 데이터 7,172장으로 구성되었다.

5.2 실험 결과

일반적인 PC 환경에서 그림 6의 CNN에 대하여 5가지 보간법을 통해 이미지가 축소된 경우와 DCT를 이용한 경우의 이미지 분류 정확도를 표 3에 정리하였다. Original은 이미지의 축소 없이 28x28 이미지를 CNN에 적용한 결과이다. 자원이 제한되지 않은 PC 환경의 실험결과 이미지를 축소하지 않은 경우나 축소한 경우, 제안된 DCT 방법의 성능 차이는 크지 않지만, Area 방식과 DCT 방식이 Original과 가장 유사한 성능을 보인 것을 알 수 있다. 이를 토대로 보간법 중 가장 정확도가 높은 Area 기법과 제안한 DCT 기법을 뉴로모픽 아키텍처 기반 FPGA 임베디드 보드에서 정확도에 대한 실험을 진행하였다. 또한, 이 결과에서 보간법의 경우 이미지를 축소시키는 과정에서 특징점이 균일하게 소실되었지만, DCT의 경우 이미지 정보를 압축 후 고주파 정보를 삭제하여 이미지의 전반적인 특성을 보존하기 때문인 것으로 판단한다. 이들 실험 결과를 바탕으로 Area 보간법과 DCT를 이용한 특징 표현을 이용하여 자원이 제한된 두 개의 FPGA 뉴로모픽 임베디드 보드에서 실험을 진행하였다.

표 3. PC 환경에서 특징 표현에 따른 CNN 기반 이미지 분류 비교
Table 3. Comparison of CNN based image classifications according to feature representation on PC environment

MNIST Datasets	Original (%)	Nearest (%)	B-Linear (%)	Area (%)	Bicubic (%)	Lanczos (%)	DCT (%)
Digits	99.44	98.68	99.08	99.19	99.07	99.01	99.25
Fashion	90.07	88.17	88.22	88.95	88.26	87.97	89.16
Sign Language	97.29	93.34	94.76	95.71	95.07	92.82	96.91

표 4. DE1-SoC 디바이스 성능 비교
Table 4. Performance comparison of DE1-SoC device

MNIST Datasets	Original (%)	Area (%)	DCT (%)
Digits	33.59	76.52	83.91
Fashion	44.76	69.10	70.51
Sign Language	61.45	95.83	96.80

FPGA 보드는 연산할 수 있는 뉴런의 개수가 제한됨에 따라 원본 이미지의 정확도를 유지하는 경우 연산에 사용되는 뉴런의 수가 줄어들어 DE1-SoC 보드에서 각각 33.59%, 44.76%, 61.45%으로 CNN 실험에 비해 낮은 정확도를 갖는 것을 확인하였다. 또한, 모델링 할 수 있는 뉴런의 수가 2배인 PYNQ 보드에서도 유사하게 성능이 저하됨을 확인할 수 있었다.총 16K를 지원하는 DE1-SoC 보드에서 28x28 원본 이미지 입력에 대한 은닉층은 16K / 784의 값인 20개의 은닉 뉴런만을 사용하여 모델링 할 수 있다. 따라서 이미지의 해상도를 14x14로 낮추면 은닉 뉴런 수를 81개로 늘릴 수 있다. Area 보간법과 DCT 방법을 사용하여 해상도를 낮추고 비교한 결과, CNN 실험과 마찬가지로 DCT 기법을 사용한 경우가 Area 보간법에 비해 더 좋은 정확도를 보였다.

표 5. PYNQ 디바이스 성능 비교
Table 5. Performance comparison of PYNQ device

MNIST Datasets	Original (%)	Area (%)	DCT (%)
Digits	76.83	86.91	91.83
Fashion	65.96	76.41	77.55
Sign Language	82.85	98.02	98.11

다음으로는 표 5와 같이 PYNQ 보드를 사용하여 실험을 진행하였다. PYNQ 보드의 자원은 표 1과 같이 사용할 수 있는 총 뉴런의 수는 32K개의 뉴런을 사용할 수 있다. 실험 결과 원본 이미지에 대하여 각각 76.83%, 65.96%, 82.85%의 정확도를 보여 뉴런의 수가 적은 DE1-SoC보다는 성능이 향상된 것을 알 수 있었다.

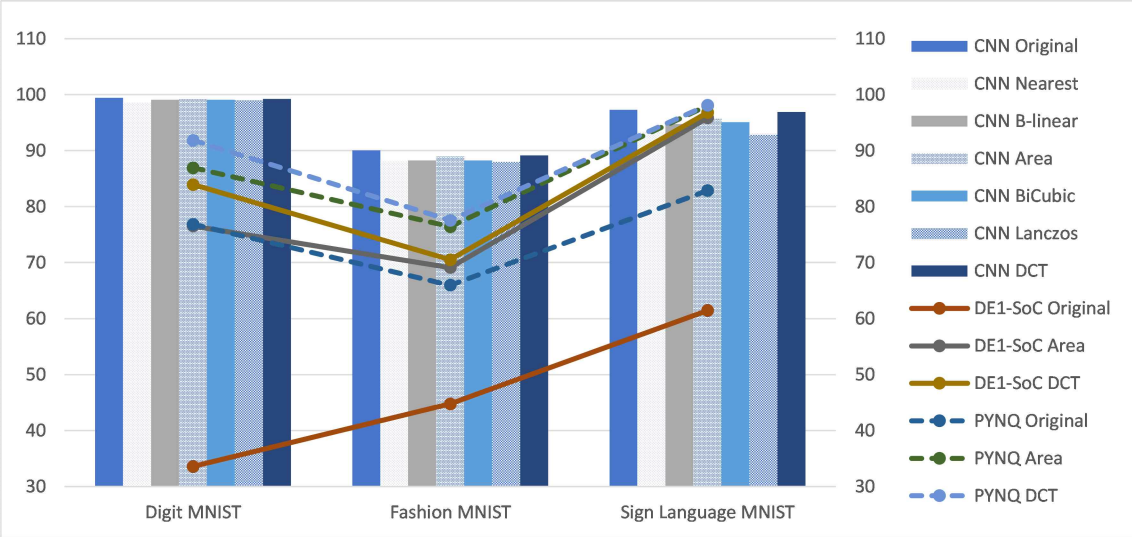


그림 9. 장치(환경)별 분류 성능 비교
Fig. 9. Comparison of classification performance by device

PC 환경과 FPGA 임베디드 보드에 대한 전체 분류 성능을 그림 9에 정리하였다. PC 환경에서나 두 FPGA 임베디드 보드에서 모두 보간법을 사용하는 것보다 DCT를 사용하였을 때 정확도가 더 향상됨을 보여, 제안된 방법이 유효함을 알 수 있었다. PC 환경의 실험과 달리 두 임베디드 보드 모두 수화 인식 실험에서 다른 실험에 비하여 높은 성능을 보이고 있으며, 특히 PYNQ의 경우 PC 환경의 Original 실험 결과보다 좋은 결과를 보였다. 이는 이미지 축소나 압축과정에서 특징 분포가 변경되고 소실된 데이터가 오히려 성능 향상을 보였을 것으로 판단한다. 특징 분포의 변화와 소실된 데이터에 대한 분석은 향후 연구에서 성능 향상의 단초를 제공할 수 있을 것으로 판단한다.

6. 결 론

최근 급격한 발전을 보이는 인공지능 기술은 높은 성능임에도 불구하고 빅데이터와 높은 전력 소비량을 보이는 약점으로 IoT 환경과 같이 매우 제한된 자원에서 적용하기에 많은 어려움이 따른다. 이런 문제점을 해결하기 위하여 뉴로모픽 구조 기반 신경망 기술들이 제안되었으며, 본 논문에서는 제한된 자원에서 성능의 저하를 최소화하여 임베디드 환경에서 사용할 수 있도록 DCT 방식을 이용하여 이미지를 표현하는 방법을 제안하였다. 제안된 방법의 유효성을 확인하기 위하여 PC 환경에서 CNN 구조를 기반으로 5가지의 보간법과 DCT 기법을 이용하여 이미지의 해상도를 낮추어 정확도를 비교하였다. 실험 결과 예측한 바와 같이 은닉 뉴런의 수가 증가하여 정확도가 향상됨을 보였고, DCT 기법을 사용한 것이 보간법을 사용하였을 때보다 성능이 향상됨을 확인하였다. 추후 연구로는 본 논문에서 제안한 DCT 특징 표현을 뉴로모픽 구조기반의 SNN으로 모델링하여 자원이 제한된 환경에서 성능 향상을 높이고자 한다. 아울러 DCT 기반의 특징 표현과 이에 특화된 SNN 모델이 성능을 향상시킨다면, 임베디드 보드에 최적화되고 경량화된 DCT 기법의 개선 등에 대한 연구도 필요할 것으로 보인다.

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2019-0-00708, 뉴로모픽 아키텍처 기반 자율형 IoT 응용 통합개발환경).

참 고 문 헌

[1] Another Way of Looking at Lee Sedol vs AlphaGo, Sep. 2021. https://jacquesmattheij.com/another-way-of-looking-at-lee-sedol-vs-alphago/,

[2] A. So, “Brain neural circuit neuromorphic chip”, Convergence Research Policy Center Convergence Weekly TIP, Vol. 104, pp. 2-3, Jan. 15, 2018. https://crpc.kist.re.kr/common/attachfile/attachfileDownload.do?attachNo=00004268

[3] S. Kim, Y.-S. Yun, J. Jung, “Design of a Framework for supporting Neuromorphic Hardware in IoT platform”, Journal of Information Science Vol. 38, No. 2 pp. 51-57, Feb. 2020. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09304619

[4] J. Jeong, S. Kim, J. Kim, B. Kim, J. Jung, “Development of the IoT Application Generation Automation Tool for supporting Neuromorphic Hardware”, Proceedings of the Korean Society of Information Sciences Conference, pp. 819-821, Dec. 2020. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10529799

[5] J. Wang, H. Wang, X. Zhu, P. Zhou, “A Deep Learning Approach in the DCT Domain to Detect the Source of HDR Images”, Electronics 2020, 9(12), 2053 Nov. 29, 2020 DOI: https://doi.org/10.3390/electronics9122053

[6] K.-C. Shin, “The Study of Comparison of DCT-based H.263 Quantizer for Computative Quantity Reduction”, Journal of Convergence Signal Processing Society Vol. 9, No. 3, pp. 195 - 200, July. 30, 2008. https://www.koreascience.or.kr/article/JAKO200828939697904.page

[7] Y.-H. Moon, “An Efficient DCT Calaculation Method Based on SAD”, Journal of the Korean Telecommunications Society Vol. 28, No. 6C, pp. 602-608 Jun. 1, 2003. https://www.koreascience.or.kr/article/JAKO200311921706813.page

[8] W. Guo, H. E. Yantir, et al. “Towards Efficient Neuromorphic Hardware: Unsupervised Adaptive Neuron Pruning”, Electronics 2020, vol. 9, no. 7, 1059, DOI: https://doi.org/10.3390/electronics9071059

[9] Nengo: a Python tool for building large-scale functional brain models. https://www.frontiersin.org/articles/10.3389/fninf.2013.00048/full, Sep. 2021

[10] R. Wang, C. S. Thakur, G. Cohen, T. J. Hamilton, J. Tapson, A. v. Schaik, “Neuromorphic hardware architecture using the neural engineering framework for pattern recognition”, Indian Institution Of Industrial Engineering transactions on biomedical circuits and systems, Vol. 11, No. 3, pp. 574-584, June. 3, 2017. DOI: https://doi.org/10.1109/TBCAS.2017.2666883

[11] O. Rukundo, H. Cao, “Nearest Neighbor Value Interpolation”, International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 3, No. 4, pp. 1-6, Nov. 8, 2012. DOI: https://doi.org/10.14569/IJACSA.2012.030405

[12] P. R. Smith, “Bilinear interpolation of Digitsal images”, Ultramicroscopy, Vol. 6, No. 2, pp. 201-204, Jan. 5, 1981. DOI: https://doi.org/10.1016/0304-3991(81)90061-9

[13] R. G. Keys, “Cubic convolution interpolation for Digitsal image processing”, Indian Institution Of Industrial Engineering Transactions on Acoustic, Speech, and Signal Processing, Vol. 29, No. 6, pp. 1153-1160, Dec. 6, 1981. DOI: https://doi.org/10.1109/TASSP.1981.1163711

[14] B. N. Madhukar, R. Narendra, “Lanczos resampling for the Digitsal processing of remotely sensed images”, In Proceedings of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking, pp.403-411, July. 10, 2013. DOI: https://doi.org/10.1007/978-81-322-1524-0_48

[15] N. Ahmed, T. Natarajan, K.R. Rao “Discrete cosine transform”, Indian Institution Of Industrial Engineering Transactions on Computers 23, pp. 90 - 93, Jan. 1, 1974. DOI: https://doi.org/10.1109/T-C.1974.223784

[16] Myunggeun Ji, Junchul Chun, Namgi Kim, “An Improved Image Classification Using Batch Normalization and CNN”, Journal of Internet Computing and Services, Vol. 19, No. 3, pp. 35 - 42, Dec. 26, 2018. DOI: https://doi.org/10.7472/jksii.2018.19.3.35

저 자 소 개

정재혁(Jae-Hyeok Jeong)

2020.2 한남대학교 정보통신공학과 졸업
2020.3-현재 한남대학교 정보통신공학과 석사과정
<주관심분야> 인공지능, 이미지 전처리

정진만(Jinman Jung)

2008.02 서울대학교 컴퓨터공학과 졸업
2014.02 서울대학교 컴퓨터공학과 박사
2014.9-2021.2 한남대학교 정보통신공학과 조교수
2021.3-현재 인하대학교 컴퓨터공학과 부교수
<주관심분야> 운영체제, 임베디드 시스템,IoT

윤영선(Young-Sun Yun)

1990.2 KAIST 전산학과 졸업
1992.2 KAIST 전산학과 석사
2001.2 KAIST 전산학과 박사
2006.4-2007.2 한국전자통신연구원 초빙연구원
2012.8-2013.7 University of Washington 방문학자
2001.3-현재 : 한남대학교 교수
<주관심분야> 음성인식, 음성변환, 화자인식, 인공지능, 내장형시스템, 저작권침해, 유사도, 완성도 감정, 오픈소스 등