부동소수점 형식에 따른 스파이킹 신경망의 학습 및 정확도 분석

곽명진, 서효주, 김용태*
경북대학교 컴퓨터학부

*yongtae@knu.ac.kr

Training and Accuracy Analysis of Spiking Neural Network using Floating-Point Formats

Myeongjin Kwak, Hyoju Seo, and Yongtae Kim*
School of Computer Science and Engineering, Kyungpook National University

요 약

본 논문은 에너지 효율적인 스파이킹 신경망(spiking neural network) 구현을 위해 다양한 부동소수점 형식(floating-point format)의 정밀도에 따른 학습의 정확도를 체계적으로 분석한다. IEEE 754 기반의 32 비트 단정밀도 부동소수점 및 비트폭이 축소된 부동소수점 형식을 스파이킹 신경망의 Leaky Integrate-and-Fire (LIF) 뉴런에 적용하여 축소된 부동소수점 형식이 스파이킹 신경망에 미치는 영향을 분석한다. 손글씨 이미지 MNIST 데이터를 분류하기 위한 두 개의 레이어로 구성된 스파이킹 신경망에 다양한 정밀도의 부동소수점 형식을 적용하여 비지도 학습을 수행하고 결과를 분석한다. 지수비트가 4 비트 이상인 16 비트폭의 부동소수점 형식은 32 비트 단정밀도 부동소수점 형식과 유사한 정확도로 학습할 수 있다. 또한, 신경망의 학습을 위해서는 최소 지수 4 비트와 가수 6 비트가 필요하며 이를 하드웨어로 구현할 경우 다른 부동소수점 형식에 비해 면적, 전력, 및 에너지 효율이 우수하다.

Ⅰ. 서 론

인공 신경망(artificial neural network) 발전으로 다양한 분야에 기계 학습(Machine Learning) 기술이 적용되고 있다. 이를 처리해야 할 데이터도 급격하게 증가하였고 필요한 컴퓨팅 자원과 전력/에너지 소모도 지속적으로 증가하고 있다. 이러한 문제를 해결하기 위한 방법 중의 하나로 뉴로모픽 컴퓨팅(Neuromorphic Computing) 기술이 있다. 이는 포유류의 뇌를 모방하는 기술로 스파이킹 신경망(spiking neural network)으로 구현할 수 있다. 생물학적 뇌는 데이터 처리에 대하여 매우 우수한 전력 및 에너지 효율성을 보여주고 이러한 신경구조를 활용한 뉴로모픽 컴퓨팅 기술은 전력/에너지 효율적인 컴퓨팅을 위한 해결책이 될 수 있다 [1].

스파이킹 신경망의 뉴런은 Hodgkin-Huxley, Izhikevich, 그리고 Leaky-Integrated and Fire (LIF) 모델 등을 사용하여 구현할 수 있다. 이러한 뉴런 모델을 적용한 신경망은 막 전위(membrane potential)가 문턱 전압(threshold voltage)보다 높을 때 스파이크(Spike)로 반응이 전달되는 이벤트 기반 방식이기 때문에 높은 전력 효율을 가진다. 이는 문자 인식, 컴퓨터 비전, 그리고 음성인식 등 다양한 분야에서 사용될 수 있다.

스파이킹 신경망은 GPU (Graphic Processing Unit) 등을 사용하여 32-비트 단정밀도부동소수점 형식(single precision floating-point format)의 연산을 수행하여 동작한다 [2]. 이때 부동소수점 연산과 표현을 위해 많은 하드웨어 자원이 필요하다. 32-비트 단정말도 부동소수점보다 낮은 정밀도를 가지는 부동소수점 형식과 연산으로 학습이 가능하다면 하드웨어 자원을 크게 절감할 수 있다. 따라서 스파이킹 신경망의 전력/에너지 효율성을 위해 다양한 정밀도의 부동소수점 연산을 스파이킹 신경망에 적용하여 성능을 평가하는 것은 매우 중요하다.

본 연구에서는 전력/에너지 효율적인 스파이킹 신경망 프로세서를 구현을 위해 다양한 정밀도의 부동소수점 형식을 스파이킹 신경망에 적용하고 정확도에 미치는 영향을 분석한다. LIF 뉴런 모델을 사용한 스파이킹 신경망에서 대표적인 비지도 학습 방법인 STDP (Spike Timing Dependent Plasticity)를 사용하여 손글씨 이미지 MNIST 데이터를 학습하였다. 실험 결과로 학습을 위해 적어도 4-비트 지수(exponent)와 6-비트 가수(mantissa)가 필요한 것을 확인하였다.

Ⅱ. 부동 소수점 형식

현대의 컴퓨터에서 사용되는 부동소수점 형식은 IEEE-754 에 정의되어 있으며, 부호, 지수, 그리고 가수 부분으로 구성된다. 단정밀도 부동소수점(FP32)은 부호, 지수, 그리고 가수가 각각 1-비트, 8-비트, 그리고 23-비트로 구성된다 [3]. 표 1 은 다양한 부동소수점 형식과 각 부분의 비트 수를 나열한다. Tensorflow-32 (TF32) 형식과 Brain floating-point (BF16) 형식은 FP32 와 지수 비트의 크기는 같으나 가수 비트는 각각 10-비트 그리고 7-비트로 구성된다 [4]. 반정밀도 부동소수점(FP16)과 DLFloat(DLF16) 형식은 지수와 가수부분에서 1 비트의 차이가 있다 [5]. 또한, EFloat16 Average (EF16_A) 형식은 지수 3-비트, 가수 12-비트로 구성되며, 8-비트 부동소수점 형식인 Minifloat(FP8)은 지수 4-비트, 가수 3-비트로 구성된다 [6].

표 1. 비트폭에 따른 다양한 부동 소수점 형식.

Format	*sign*	*exponent*	*mantissa*	*total*
FP32	1	8	23	32
TF32	1	8	10	19
FP16	1	5	10	16
BF16	1	8	7	16
DLF16	1	6	9	16
EF16_A	1	3	12	16
FP8	1	4	3	8

Ⅲ. 실험 결과

다양한 부동소수점 형식을 적용하기 위한 학습 모델로 [7]에서 제시된 스파이킹 신경망의 구조를 사용하였다. 첫번째 레이어는 입력 레이어로 784 개의 뉴런으로 구성되며, 두번째 레이어는 흥분성(excitatory) 뉴런과 억제성(inhibitory) 뉴런으로 구성되며 각각의 뉴런들의 개수는 동일하다. 두번째 레이어는 400 개와 1600 개의 흥분성 뉴런과 억제성 뉴런으로 구성하여 실험하였다. LIF 뉴런 모델을 사용하였고 해당 연산에 다양한 정밀도의 부동소수점 형식을 적용하였다. 학습은 훈련 데이터 6 만개와 테스트 데이터 1 만개를 사용하여 진행하였다.

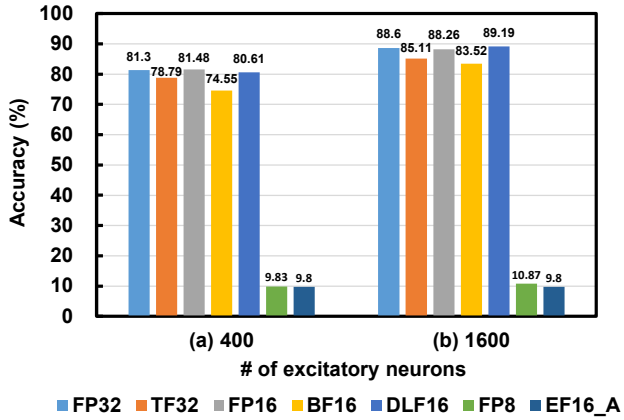

그림 1. 부동소수점 형식에 따른 스파이킹 신경망의 정확도; (a) 흥분성 뉴런 400 개, (b) 흥분성 뉴런 1600 개 일 때 정확도.

그림 1 은 다양한 정밀도를 가지는 부동소수점 형식을 적용하여 학습한 스파이킹 신경망의 정확도를 보여준다. FP32 는 뉴런이 400 개, 1600 개 일 때 각각 81.3%, 88.6%의 정확도를 가진다. FP8 과 EF16_A 를 제외한 부동소수점 형식에서는 FP32 와 비슷한 정확도를 가진다. TF32, FP16, BF16, 그리고 DLF16 을 적용한 뉴런이 400 개인 스파이킹 신경망에서 각각 78.8%, 81.5%, 74.6%, 그리고 80.6%의 정확도를 가지며, 1600 개 일 때 각각 85.1%, 88.3%, 83.5%, 89.2%의 정확도를 가진다. FP8 과 EF16_A 를 적용하였을 경우의 정확도는 뉴런이 400 개 일 때 모두 9.8%, 1600 개 일 때 각각 10.9%, 9.8%로 정확도가 크게 감소하여, 정상적으로 손글씨 분류가 되지 않는 것을 확인할 수 있다.

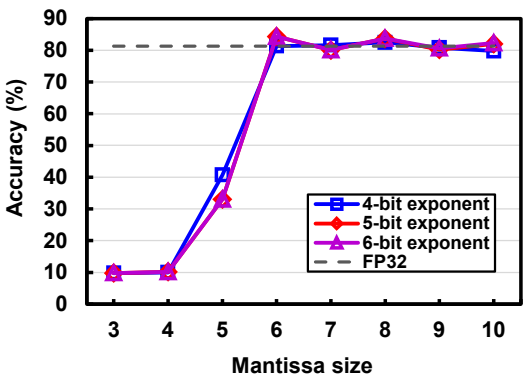

그림 2. 부동소수점 비트폭에 따른 스파이킹 신경망의 정확도.

스파이킹 신경망의 학습 정확도를 유지하기 위한 최소의 지수와 가수 비트 크기를 결정하기 위해 지수의 크기를 4, 5, 그리고 6 비트로 고정하고 가수의 크기를 변경하며 학습의 정확도를 측정하였다. 그림 2 는 흥분성 뉴런이 400 개로 구성된 스파이킹 신경망에서 지수와 가수의 크기에 따른 정확도를 보여준다. 지수가 4-비트 이상일 때 가수가 6-비트 이상이면 FP32 형식과 유사한 정확도를 보인다. 하지만, 가수가 5-비트 이하 일 때 정확도가 급격하게 감소함을 확인할 수 있다. 따라서, LIF 뉴런을 적용한 스파이킹 신경망에서 정확도의 영향 없이 학습하기 위해서 최소 4-비트 지수와 6-비트 가수가 필요한 것을 확인할 수 있다.

표 2. 다양한 부동소수점 형식의 하드웨어 비용.

Format	*Area (μm^2)*	*Power (μW)*	*Energy (fJ)*
FP32	310.08	95.43	257.76
TF32	193.60	58.31	79.58
FP16	162.88	48.32	67.27
BF16	166.72	49.75	55.15
DLF16	164.80	49.32	61.57
FP11	88.96	25.10	16.58

실험에서 적용한 다양한 부동소수점 형식의 하드웨어 비용을 산출하고 비교하기 위해 부동소수점 연산의 핵심 부분인 지수 감산기와 가수 가산기를 Verilog HDL 로 설계하여 *65-nm* CMOS 공정으로 합성하였다. 표 2 는 구현 결과이며, 학습에 실패한 형식은 제외하였고, FP11 은 지수 4-비트, 가수 6-비트를 가지는 형식이다. FP32 는 면적, 파워, 및 에너지를 가장 많이 소모하며. FP16, BF16, 그리고 DLF16 은 FP32 와 비교하면 면적과 파워는 거의 절반을 소비하고 에너지 효율은 4 배 우수하다. FP11 은 면적, 파워, 및 에너지 측면에서 가장 효율적이고, 구체적으로 FP32 와 비교하여 각각 71.3%, 73.7%, 93.6% 우수함을 확인할 수 있다.

Ⅳ. 결론

스파이킹 신경망의 LIF 뉴런에 적용가능한 다양한 정밀도를 가지는 부동소수점 형식을 분석하였다. TF32, FP16, BF16, 그리고 DLF16 형식을 스파이킹 신경망의 LIF 뉴런에 적용하여 정확도의 감소 없이 적용 가능함을 확인했다. 또한 LIF 뉴런에 적용가능한 최적 비트 크기를 분석한 결과 지수 비트가 4 비트 이상이고 가수 비트가 6 비트 이상일 때 정확도의 영향 없이 학습이 가능하다. 따라서 FP32 보다 작은 비트로 학습이 가능하고 이를 통해 하드웨어 리소스 사용량을 감소시킬 수 있다.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded bythe Ministry of Education (NRF-2019R1I1A3A01061266).

참 고 문 헌

[1] Y. Kim et al., “A Reconfigurable Digital Neuromorphic Processor with Memristive Synaptic Crossbar for Cognitive Computing,” *ACM J. Emerg. Tech. Comput. Syst.*, 2015

[2] Y. Wang et al., “Efficient Spiking Neural Network Training and Inference with Reduced Precision Memory and Computing,” *IET Comput. Dig. Tech.*, 2019.

[3] “IEEE Standard for Floating-Point Arithmetic,” IEEE Std. 754-2008, 2008.

[4] N. Burgess et al., “Bfloat16 Processing for Neural Networks,” *IEEE Symp. Comput. Arith. (ARITH)*, 2019.

[5] A. Agrawal et al., “DLFloat: A 16-b Floating Point Formatn Designed for Deep Learning Training and Inference,” *IEEE Symp. Comput. Arith. (ARITH)*, 2019.

[6] R. Bordawekar et al., “EFloat: Entropy-coded Floating Point Format for Deep Learning,” arXiv:2102.02705, 2021

[7] P. U. Diehl et al., “Unsupervised Learning of Digit Recognition using Spike-Timing-Dependent Plasticity,” Front. Comput. Neurosci., 2015.