시냅스 소자 구현을 위한 균일 양자화 방식

이재은, 이철준, 이대석, 김동욱, 서영호
광운대학교
jelee @kw.ac.kr

Uniform Quantization Method for Synaptic Device

Jae Eun Lee, Chul Jun Lee, Dae Seok Lee, Dong Wook Kim, Young Ho Seo
Kwangwoon University

요 약

본 논문에서는 뉴로모픽 시스템 구현을 위해 시냅스 소자의 비선형적인 전도도를 고려한 균일 양자화 방식을 제안한다. 소프트웨어로 학습시킨 가중치에 최댓값을 나누는 것으로 정규화를 수행한다. 그 다음, 제안하는 균일 양자화 방식을 수행한다. 양자화를 수행할 때 소자의 제한적인 전도도 레벨을 고려하여 5 부터 25 레벨로 설정하여 실험하였다. 그 결과 MNIST 시험 데이터 세트의 정확도가 10 레벨에서 95.75%로, 소프트웨어의 정확도와 1%미만의 차이를 가진다.

1. 서론

인공신경망 구현을 위한 하드웨어로 뉴로모픽 소자에 대한 연구가 진행중이다[1]. 이 중에 이상적인 연구는 많이 진행되었지만 시냅스 소자의 제한적인 요소들이 고려되지 않았기 때문에 하드웨어 구현가능성은 거의 없다. 구현가능성을 향상시키기 위해서는 소자의 특성을 고려한 비이상적인 연구가 많이 진행되어야 한다.

따라서 본 논문에서는, 시냅스 소자 구현을 위한 비선형적인 전도도를 고려하여 균일 양자화 방식을 제안한다. 양자화 레벨도 제한적인 전도도 레벨을 고려하여 5 부터 25 로 설정하여 실험하였다. 실험 결과를 기존의 방식, 소프트웨어 결과와 비교 및 분석한다.

2. 제안하는 방식

데이터 세트는 28×28 사이즈의 영상인 MNIST 를 사용하였다. 60,000 장의 학습 데이터와 10,000 장의 시험 데이터로 이루어져있다. 학습 데이터로 학습을 진행한 후 시험 데이터로 네트워크를 평가한다.

네트워크 구조는 그림 1 에 나타내었다. 784(28*28) 크기의 입력 노드, 128 크기의 히든 노드, 그리고 10 크기의 출력 노드로 구성하였다. 활성화 함수는 ReLU (Rectified Linear Unit) , 학습률(learning rate) 은 0.5, 그리고 에폭(epoch)은 200 으로 설정하였다.

소프트웨어로 학습시킨 가중치에 대하여 정규화를 수행한 다음, 제안하는 양자화 방식을 수행하고 시험 데이터 세트의 정확도를 측정한다. 정규화는 가중치와 전도도의 크기를 맞춰주기 위해 수행한다. 가중치에 최댓값을 나누는 것으로 수행한다[2]. 제안하는 양자화 방식을 수행하기 위해서는 먼저, 특정 레벨에 따른 균일 양자화를 수행한다. 균일 양자화를 수행한 가중치 값을 거리가 가장 가까운 참고 값으로 변형시킨다. 변형시킨 가중치로 시험 데이터 세트의 정확도를 측정한다. 참고 값은 시냅스 소자인 PCMO 의 모델링 함수를 50 레벨로 샘플링한 값으로, 소자가 실제로 가질 수 있는 전도도를 고려한 값이다.

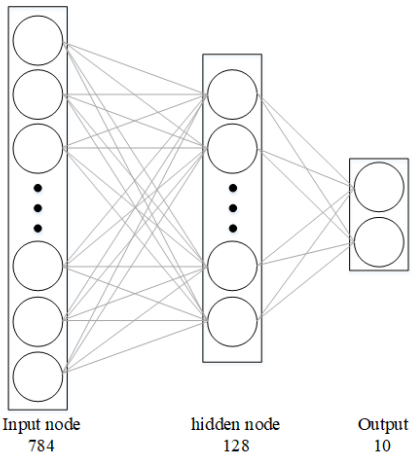

그림 1. 네트워크 구조

Figure 1. Network structure

3. 실험 결과

소자의 특성을 그대로 적용한 기존의 방식과 제안하는 방식을 비교하기 위해 레벨 변화에 따른 MNIST 시험 데이터 세트 정확도를 그림 2 에 나타내었다. 실험한 레벨의 모든 결과에 대해 제안하는 방식이 기존 방식에 비해 높은 정확도를 가지는 것을 볼 수 있다. 레벨이 감소하면서 두 방식의 정확도 차이가 증가하고 있으며, 특히 5 레벨의 경우에는 11% 정도의 정확도 차이를 가진다. 뿐만 아니라, 제안하는 방식은 10 레벨에서 95.75%를 가지는데, 이는 소프트웨어의 정확도인 96.47%와 1% 미만의 차를 보인다.

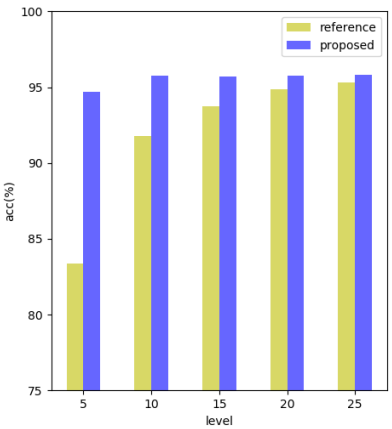

그림 2. 기존의 방식과 제안하는 방식의 레벨 변화에 대한 정확도

Figure 2. Accuracy according to level change of sampling method and the proposed method

4. 결론

본 논문에서 시냅스 소자가 가지는 전도도 값과 소자가 현실적으로 표현할 수 있는 레벨을 고려한 균일 양자화 방식을 제안하였다. 이는 이상적인 방식이 아니라 비이상적인 방식으로, 하드웨어 구현가능성을 향상시켰을 뿐만 아니라 소프트웨어의 정확도와 비교했을 때 1% 미만의 차를 가진다.

감사의 글

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF 2018R1D1A1B07043220).

참고문헌

[1] A. Thomas, “Memri stor-based neural networks ”, J. Phys. D Appl. Phys., vol. 46, no. 9, 2013.

[2] C. Zhu, S.Han, H. Mao, and W. J. Dally, “Trained ternary quantization, ” in Proc. ICLR., 2017.