비트 효율성 향상을 위한 학습 기반 특징값 크기 제한 활성화 함수

서정윤, 이하림*
금오공과대학교
seojeongyun@kumoh.ac.kr, *hrlee@kumoh.ac.kr

A Learnable Activation Function to Limit Feature Values for Bit Efficiency

Seo Jeongyun, Lee Harim*
Kumoh National Institute of Technology

요 약

최근 인공지능의 실생활 적용을 위해 Neural Processing Unit(NPU) 중 하나인 AI 하드웨어 가속기를 활용한 딥러닝 모델 추론 연구가 대두되고 있다. 다만, 딥러닝 네트워크의 학습은 서버 및 데이터 센터와 같이 상대적으로 풍부한 자원이 확보되어 있는 환경에서 진행되는 것과 달리, 실생활 적용을 위한 추론용 NPU의 경우 한정된 자원으로 딥러닝 네트워크의 추론을 구동해야 하기에 메모리 및 전력 자원을 효율적으로 활용해야만 한다. 그 결과 제한된 자원 환경에서의 딥러닝 모델 추론 효율성 관련 연구가 꾸준히 진행되어 오고 있다. 최근 딥러닝 네트워크의 특징맵 값의 비트를 제한하여 메모리 효율성을 극대화하는 연구가 있으며, 이에 본 논문에서는 학습 가능 특징값 제한 활성화 함수를 개발한다. 제안하는 활성화 함수의 경우, 학습 가능 임계치 변수를 도입하여 출력 활성화 값의 상한을 학습 도중에 제한함으로써 각 계층별 특징맵 값이 최적으로 제한될 수 있다. 해당 활성화 함수를 활용하여 순전파 단계에서 출력 값의 상한을 학습에 적응적으로 제한하고 역전파 단계에서는 상한값에 대한 편미분 값을 계산하여 학습 과정에서 상한값이 네트워크 구조에 적응적으로 최적화될 수 있도록 했다. 해당 활성화 함수를 완전 연결층 네트워크 학습에 적용하고 기존 활성화 함수들 대비 성능을 비교하는 실험을 진행하였다. 해당 실험에서 제안하는 함수가 특징맵의 표현력을 손상시키지 않으면서 상한 비트를 학습시 동적으로 제한할 수 있음을 확인하였다.

Ⅰ. 서 론

최근 다양한 산업 분야 및 실생활에 인공지능을 접목시키기 위해 AI 하드웨어 가속기에 대한 연구 개발이 활발히 진행되고 있다. 다만, AI 하드웨어 가속기는 한정된 자원으로 딥러닝 네트워크를 구동해야 한다는 점에서 메모리 및 전력 자원을 효율적으로 활용해야 할 필요가 있으며, 이에 따라 제한된 자원 환경에서의 딥러닝 모델 추론 효율성에 대한 관심이 크게 증가하고 있다.

특히, AI 하드웨어 가속기의 메모리 사양은 소프트웨어 상에서 학습이 완료된 딥러닝 네트워크의 가중치 및 특징맵 값에 의존적이라는 점에서 딥러닝 네트워크의 가중치 및 특징맵 값의 비트를 효과적으로 제한해야 저사양의 메모리만으로도 AI 하드웨어 가속기 상에서 효율적인 딥러닝 모델 추론이 가능하다. 이에 성능을 유지할 수 있으면서 가중치 및 특징맵 값의 비트를 최소화하기 위한 연구가 되고 있다 [1-4]. 하지만 기존 연구들의 경우, 딥러닝 네트워크의 학습이 완료된 후 특징값을 제한하는 최댓값을 통계적으로 설정하기 때문에 네트워크의 성능 저하가 반드시 발생한다는 한계점을 가지고 있다.

본 논문에서 제안하는 학습 가능 특징값 제한 활성화 함수에서는 특징값의 최댓값을 제한하는 학습 변수 W_{act}가 네트워크 성능은 유지하면서 비트 수를 최소화할 수 있는 방향으로 학습된다. 따라서, 기존 연구와는 달리 학습 시에 딥러닝 네트워크의 표현력은 최대한 보존하면서 모델의 계층 별로 최적의 비트 제한값을 학습할 수 있기 때문에 기존 연구의 한계점을 개선한다.

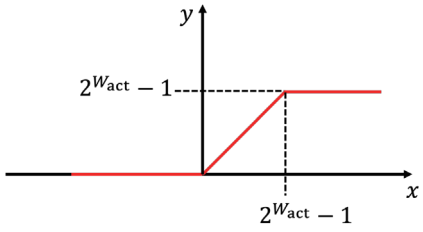

그림 1. 학습 가능 특징값 제한 활성화 함수.

Ⅱ. 학습 가능 특징값 제한 활성화 함수

기본적으로 자주 활용되는 ReLU, LeakyReLU, GELU 활성화 함수 등은 특징값에 대한 임계값를 갖지 않기 때문에, 계층 간 전달되는 특징맵의 값이 상한 없이 증가할 수 있다. 이러한 경우 하드웨어에서 설정된 메모리 규격보다 큰 비트 폭을 요구하는 특징값이 발생하게 됐을 때 해당 값을 절삭해야만 하기 때문에 네트워크 성능 저하가 발생할 수 밖에 없다. 따라서 기존 연구들은 제한된 하드웨어 규격 및 활용하는 고정소수점의 비트 수를 고려하여 통계적 관점에서 특징값을 제한하는 최댓값을 선정하였다. 하지만 이러한 경우에도 모든 특징값을 하드웨어에서 고려하는 것이 아니기 때문에 성능 저하는 여전히 발생한다.

본 연구에서는 그림 1에서와 같은 학습 가능 특징값 제한 활성화 함수를 제안한다. 특징값 제한 활성화 함수 연구의 시작으로, 기존 ReLU 활성화 함수에 특징값을 제한하는 최댓값인 $2^{W_{\text{act}}}-1$을 추가하여 특징값을 제

한하는 활성화 함수를 설계한다. 학습 과정에서 해당 활성화 함수의 W_{act} 변수가 네트워크 성능 및 비트 수 최소화를 동시에 고려하여 학습된다.

제안하는 활성화 함수를 기존 딥러닝 프레임워크에서 사용하기 위해 파이토치의 자동 미분 엔진(AutoGrad)을 기반으로 순전파 및 역전파에 대한 커스텀 연산 함수를 설계하였다. 제안하는 함수는 순전파 단계에서 출력 값의 상한을 제한하여 특징맵의 값 범위를 안정화시키고 역전파 단계에서는 상한값에 대한 편미분 값을 계산하여 학습 과정에 포함시킴으로써 네트워크 구조에 적응적으로 최적화될 수 있도록 한다.

이 방식은 특징맵의 값이 비정상적으로 커지는 것을 방지함으로써 메모리 비트 폭의 최소화를 가능하게 하고, 신경망 계층마다 서로 다른 값으로 최적화할 수 있기 때문에 계층별 유연한 비트 설정이 가능하다. 또한 학습 시에 해당 최댓값이 정해지기 때문에 모델의 표현력을 보장할 수 있다는 점에서 모델의 성능을 유지할 수 있다.

2.1 제안 특징값 제한 활성화 함수 및 역전파 수식

$$f(x;W_{act})=\begin{cases}0, & x\le 0,\\ x, & 0<x\le 2^{W_{act}}-1,\\ 2^{W_{act}}-1, & 2^{W_{act}}-1<x.\end{cases} \quad (1)$$

$$\frac{\partial f(x;W_{act})}{\partial x}=\begin{cases}0, & x\le 0,\\ 1, & 0<x\le 2^{W_{act}}-1,\\ 0, & 2^{W_{act}}-1<x.\end{cases} \quad (2)$$

$$\frac{\partial f(x;W_{act})}{\partial W_{act}}=\begin{cases}0, & x\le 0,\\ 0, & 0<x\le 2^{W_{act}}-1,\\ 2^{W_{act}}\ln 2, & 2^{W_{act}}-1<x.\end{cases} \quad (3)$$

제안하는 활성화 함수는 수식(1)과 같이 최댓값이 $2^{W_{act}-1}$로 제한되는 일차 함수이며, 수식(1)은 자동 미분 엔진 기반 커스텀 연산 함수의 순전파 구현에서 사용되었다.

수식(2)는 수식(1)을 x에 대해 미분한 것이고 수식(3)은 동일한 수식을 W_{act}로 미분한 것이다. 두 수식은 커스텀 연산 함수의 역전파 설계에 사용되었으며, 역전파 과정에서 W_{act}의 기울기가 계산되고 경사 하강법에 의해 학습된다.

본 연구에서 설계한 커스텀 연산 함수를 통해 신경망의 손실 값을 줄이는 방향으로 학습을 진행하면서 동시에 특징맵 값의 정수부 비트를 줄이는 방향으로 학습할 수 있고 이는 모델의 성능을 유지하면서도 특징맵 값을 안정화시킬 수 있도록 돕는다.

Ⅲ. 실험 및 분석

3.1 학습 조건

완전 연결층 신경망은 입력 노드 784개와 계층 별 500, 400, 200, 100, 50, 10개의 노드를 갖는 5층 신경망 구조를 사용했고 Fashion MNIST를 학습에 사용했다. 네트워크 가중치 학습을 위해 Optimizer는 SGD, lr=1e-3, weight_decay=5e-3을 사용했고 W_{act}를 학습시키기 위해 Optimizer는 SGD, lr=1e-4, weight_decay=5e-4를 사용했으며 W_{act}의 초기값은 1로 설정했다. 이는 초기 특징맵 값이 1로 제한됨을 의미한다.

3.2 실험 결과

표 1은 3.1절의 학습 조건 하에 제안하는 학습 가능 특징값 제한 활성화 함수를 각 계층에 적용했을 때, 기존 ReLU 대비 줄어든 특징맵 비트 폭과 제거된 특징맵 값의 비율을 나타낸다. 실험 결과에서 볼 수 있듯이, 1번 계층은 기존 ReLU와 동일한 비트 폭을 유지했고 이후 2번 및 3번 계층에서는 각각 1비트, 4번 및 5번 계층에서는 각각 2비트 감소하였다. 그리고

표 1. 학습 가능 특징값 제한 활성화 함수 적용으로 인한 각 계층별 기존 ReLU 대비 줄어든 특징맵 비트 폭 및 제거된 특징맵 값의 비율

	1번 계층	2번 계층	3번 계층	4번 계층	5번 계층
줄어든 비트	0 bit	1 bit	1 bit	2 bits	2 bits
제거 비율	0.8%	2.71%	3.18%	5.73%	8.75%

계층별 제거된 특징맵 값의 비율은 1번 계층에서 0.8%, 2번 계층에서 2.71%, 3번 계층에서 3.18%, 4번 계층에서 5.73%, 5번 계층에서 8.75%이다. 1번 계층에서 기존 ReLU와 동일한 비트 폭을 유지하고 제거된 값의 비율이 적은 것은 1번 계층이 입력 데이터와 가장 가까운 계층이기 때문에 해당 계층에서 비트 폭을 크게 제한할 경우 네트워크가 입력 데이터에 대한 표현력을 상실할 수 있기 때문에, 네트워크 자체적으로 1번 계층의 값을 최대한 유지하고자 한 것이라 판단된다.

또한, 학습 가능 특징값 제한 활성화 함수의 도입으로 인한 정확도 손실은 0.3%에 불과하여, 모델의 성능을 유지하면서도 각 계층의 활성화 값 분포에 따라 상한 비트를 동적으로 최적화할 수 있음을 입증하였다. 이는 학습 가능 특징값 제한 활성화 함수가 특징맵의 표현력을 손상시키지 않으면서도 비트 폭을 효과적으로 줄이는 데 기여함을 보여준다.

Ⅳ. 결론

제안하는 학습 가능 특징값 제한 활성화 함수는 활성화 값을 제한하기 위한 학습 가능 임계값을 도입하였다. 제안 활성화 함수를 신경망 학습에 사용함으로써 계층 별 특징맵 값을 표현하기 위한 비트 수를 최소화하는 방향으로 최적화할 수 있다. 이를 통해 모델의 표현력을 유지하여 성능을 보장하면서도 하드웨어 자원이 제한된 AI 하드웨어 가속기 환경 등에서 특징맵 값을 저장하는 메모리 비트 사용의 효율성을 극대화할 수 있다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) RS-2025-00557827.

참 고 문 헌

[1] M. Courbariaux, Y. Bengio, and J.-P. David, “BinaryConnect: Training Deep Neural Networks with Binary Weights during Propagations,” Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2015.

[2] P. Judd, J. Albericio, T.Hetherington, T.Aamodt, N. E. Jerger, R. Urtasun, and A.Moshovos, “Reduced-Precision Strategies for Bounded Memory in Deep Neural Nets,” CoRR, arXiv:1511:05236, 2016.

[3] P. Gysel, M. Motamedi, and S. Ghiasi, “Hardware-Oriented Approximation of Convolutional Neural Networks,” CoRR, arXiv:1604.03168, 2016.

[4] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or -1,” CoRR, arXiv:1602.02830, 2016.