(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년10월27일
(11) 등록번호 10-2318219
(24) 등록일자 2021년10월21일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/063* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/063 (2013.01)
(21) 출원번호 10-2019-0062919
(22) 출원일자 2019년05월29일
심사청구일자 2019년05월29일
(65) 공개번호 10-2020-0137122
(43) 공개일자 2020년12월09일
(56) 선행기술조사문헌
1-Bit 합성곱 신경망을 위한 정확도 향상 기법. 임성훈, 이재홍. 한국전기전자학회. 전기전자학회 논문지 22(4), 2018.12, 1115-1122(8 pages).*
이진화된 컨벌루션 신경망의 효율적인 SIMD 구현. 김성찬, 신지훈, 박용민, 김태환. 대한전자공학회. 전자공학회논문지 55(1), 2018.1, 49-56(8 pages).*
KR1020190022237 A
KR1020190070044 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
한발대학교 산학협력단
대전광역시 유성구 동서대로 125 (덕명동)
(72) 발명자
이재홍
대전광역시 유성구 엑스포로 448 엑스포아파트 305-403
임성훈
대전광역시 유성구 동서대로5번길 39-5 노블레스빌 202호
김상혁
대전광역시 대덕구 동춘당로 178 보람아파트, 105-708
(74) 대리인
정희환

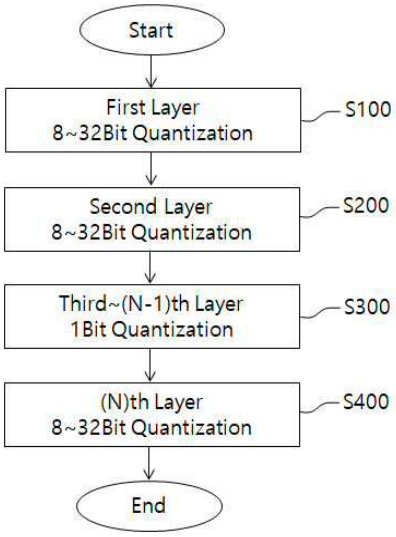
전체 청구항 수 : 총 2 항 심사관 : 박미정

(54) 발명의 명칭 **1비트 신경망 학습 방법**

(57) 요 약

본 발명은 1-Bit 신경망 학습 방법에 관한 것으로, 첫 번째 층과 두 번째 층을 8~32Bit로 계산하는 단계; 총 N개의 인공 신경망 층에서 세 번째 층부터 N-1번째의 층까지는 이진 양자화를 적용하는 단계; 및 N번째 층은 8~32Bit로 계산하는 단계;를 포함하는 것을 특징으로 한다.

대 표 도 - 도1



공지예외적용 : 있음

명 세 서

청구범위

청구항 1

컴퓨터로 첫 번째 층과 두 번째 층을 8~32Bit로 계산하는 단계;

총 N개의 인공 신경망 층에서 세 번째 층부터 N-1번째의 층까지는 이진 양자화를 적용하는 단계; 및

컴퓨터로 N번째 층은 8~32Bit로 계산하는 단계;를 포함하되,

상기 이진 양자화는 1-Bit 연산으로 하기 수학식 1에서 w_{ji}와 $z_i^{\ell-1}$가 하기 수학식 2를 통해 -1 또는 +1로 양자화된 값을 통해 연산하는 것을 특징으로 하는 1비트 신경망 학습 방법.

[수학식 1]

$$y_j^{\ell} = \sum_{i=1}^{m} w_{ji}^{\ell} \cdot z_i^{\ell-1}, (1 \le j \le n)$$

$$z_j^{\ell} = f(y_j^{\ell}), (1 \le \ell \le L)$$

$$z_i^1 = y_j^1 = x_i$$

[수학식 2]

$$x^b = sign(x), \begin{cases} +1, x \ge 0 \\ -1, x < 0 \end{cases}$$

여기서, m은 신경망의 출력층 개수, ℓ은 신경망층, L은 총 신경망층 수, w_{ji}는 신경망의 가중치, $z_i^{\ell-1}$는 이전 층의 출력값, x_i는 신경망의 입력값임.

청구항 2

삭제

청구항 3

제1항에 있어서,

상기 이진 양자화에 따른 오차를 보정하기 위해 하기 수학식과 같이 scale값을 구하여 1-Bit 연산이 적용되는 해당 층의 가중치의 절대값을 모두 더해서 개수만큼 나눈 값을 곱하여 보정하는 것을 특징으로 하는 1비트 신경망 학습 방법.

$$s^l = \frac{1}{Q \times C \times K_2 \times K_1} \sum_{q=1}^{Q} \sum_{c=1}^{C} \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} |w_{q,c,k_2,k_1}|$$

여기서, Q는 출력 채널의 개수, C는 입력 채널의 개수, K_2는 합성곱 신경망의 필터 세로 크기, K_1는 합성곱 신경망의 필터 가로 크기, $|w_{q,c,k_2,k_1}|$는 가중치의 절대값임.

발명의 설명

기 술 분 야

[0001] 본 발명은 1비트 신경망 학습 방법에 관한 것으로, 더욱 상세하게는 인공 신경망의 성능을 향상시키는 1비트 신경망 학습 방법에 관한 것이다.

배 경 기 술

[0002] 일반적으로 신경망은 32-Bit 또는 16-Bit의 부동소수점 연산을 통해 결과 값을 산출한다. 대부분의 신경망 부동소수점 연산은 MAC(Multiplier-Accumulator)이며 모델의 크기가 커질수록 연산 횟수는 크게 증가한다.

[0003] 연산의 병렬성을 증가시키기 위해 8-Bit 또는 1-Bit 연산으로 변경하는 방법이 있으며 이론적으로 32-Bit 연산보다 각각 4배 또는 32배 속도 향상을 기대할 수 있다.

[0004] 그러나, 기존의 1-Bit(XNOR-Net)의 방법은 첫 번째 층과 마지막 층만 32-Bit로 연산하고, 중간층은 모두 1-Bit로 연산하기 때문에 두 번째 층에서 특징 표현력이 떨어져 신경망의 성능이 저하되는 문제점이 있다.

발명의 내용

해결하려는 과제

[0005] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은 첫 번째 층만 유지했던 8~32Bit 연산을 두 번째 층까지 확장하여 신경망의 정확도 성능을 향상시키는 1비트 신경망 학습 방법을 제공하는 데 있다.

과제의 해결 수단

[0006] 상기와 같은 목적을 달성하기 위한 본 발명의 1-Bit 신경망 학습 방법은 첫 번째 층과 두 번째 층을 8~32Bit로 계산하는 단계; 총 N개의 인공 신경망 층에서 세 번째 층부터 N-1번째의 층까지는 이진 양자화를 적용하는 단계; 및 N번째 층은 8~32Bit로 계산하는 단계;를 포함하는 것을 특징으로 한다.

발명의 효과

[0007] 이상 같이, 본 발명에 따르면 32-Bit 신경망과 비교해서 연산 속도가 수십 배 증가하며 기존의 1-Bit 신경망보다 성능이 향상되는 장점이 있다.

도면의 간단한 설명

[0008] 도 1은 본 발명에 따른 1비트 신경망 학습 방법의 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0009] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시 예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다.

[0010] 그러면 본 발명에 따른 1비트 신경망 학습 방법의 일실시예에 대하여 자세히 설명하기로 한다.

[0011] 도 1은 본 발명에 따른 1비트 신경망 학습 방법의 순서도이다.

[0012] 도 1을 참조하면, 본 발명의 1비트 신경망 학습 방법은 먼저, 컴퓨터로 첫 번째 층을 8~32Bit로 계산한다(S100).

[0013] 이어서, 컴퓨터로 두 번째 층을 8~32Bit로 계산한다(S200).

[0014] 다음으로, 총 N개의 인공 신경망 층에서 세 번째 층부터 N-1번째의 층까지는 이진 양자화를 적용하며(S300), 컴퓨터로 N번째 층은 8~32Bit로 계산한다(S400).

[0015] 상기 8~32Bit 계산은 하기 수학식 1과 같이 나타낼 수 있다.

[0016] [수학식 1]

[0017] $$y_j^\ell = \sum_{i=1}^{m} w_{ji}^\ell \cdot z_i^{\ell-1}, (1 \le j \le n)$$

$$z_j^\ell = f(y_j^\ell), (1 \le \ell \le L)$$

$$z_i^1 = y_j^1 = x_i$$

[0018] 여기서, m은 신경망의 출력층 개수, ℓ은 신경망층, L은 총 신경망층 수, w_{ji}는 신경망의 가중치, $z_i^{\ell-1}$는 이전 층의 출력 값, x_i는 신경망의 입력값이다.

[0019] 상기 이진 양자화는 1Bit 연산을 의미하며 수학식 1에서 w_{ji}와 $z_i^{\ell-1}$가 하기 수학식 2를 통해 -1 또는 +1로 양자화된 값을 통해 수학식 1과 같이 연산을 한다.

[0020] [수학식 2]

[0021] $$x^b = sign(x), \begin{cases} +1, x \ge 0 \\ -1, x < 0 \end{cases}$$

[0022] 본 발명에서는 양자화가 적용되지 않는 첫 번째 층, 두 번째 층, 마지막 층을 제외하고는 비선형 활성화 함수 $f(y_j^\ell)$를 적용하지 않는다.

[0023] 삭제

[0024] 삭제

[0025] 또한, 본 발명에서는 하기 수학식 4, 5와 같이 신경망의 각 층의 분포가 평균이 0 분산이 1이 되도록 정규화한다.

[0026] [수학식 4]

[0027] $$mean = \frac{\sum_{i=0}^{m} x_i}{m} = 0$$

[0028] [수학식 5]

[0029] $$variance = \sqrt{\frac{\sum_{i=0}^{m} (x - mean)^2}{m}} = 1$$

[0030] 그리고, 본 발명에서는 이진 양자화에 따른 오차를 보정하기 위해 하기 수학식 6과 같이 scale값을 구하여 1-Bit 연산이 적용되는 층에 곱한다. 즉, 신경망의 해당 층의 가중치의 절대값을 모두 더해서 개수만큼 나눈 값을 곱하여 보정한다.

[0031] [수학식 6]

[0032] $$s^l = \frac{1}{Q \times C \times K_2 \times K_1} \sum_{q=1}^{Q} \sum_{c=1}^{C} \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} |w_{q,c,k_2,k_1}|$$

[0033] 여기서, 2차원 합성곱 신경망의 경우 Q는 출력 채널의 개수, C는 입력 채널의 개수, K_2는 합성곱 신경망

의 필터 세로 크기, K_1는 합성곱 신경망의 필터 가로 크기이다.

[0034] 이때, 1-Bit 양자화 신경망의 가중치를 갱신하기 위해서 양자화된 입력의 미분값인 $\frac{\partial y}{\partial x^b}$을 계산하고, 수학식 7과 같이 출력 층으로부터 계산된 미분 값을 그대로 전달한다.

[0035] [수학식 7]

[0036] $$\frac{\partial y}{\partial x^b} = \frac{\partial E}{\partial y} \cdot 1$$

[0037] 상기 수학식 3의 sign 함수의 미분은 불가능하지만 본 발명에서는 입력 값이 과 사이이면 출력 층의 미분 값을 그대로 전달하고, 그 외에는 전달하지 않는 함수를 적용한다. 이를 수학식 8과 같이 나타낼 수 있다.

[0038] [수학식 8]

[0039] $$\frac{\partial x^b}{\partial x} = \frac{\partial E}{\partial x^b} \cdot 1_{|x| \leq 1}$$

[0040] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

도면1