

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2022-0093789
(43) 공개일자 2022년07월05일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 3/04 (2006.01)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2020-0184857
(22) 출원일자 2020년12월28일
심사청구일자 2020년12월28일

(71) 출원인
한국기술교육대학교 산학협력단
충청남도 천안시 동남구 병천면 충절로 1600 (한국기술교육대학교내)
(72) 발명자
강형주
대전광역시 서구 둔산북로 160, 4동 1204호(둔산동, 한마루아파트)
김형기
충청남도 천안시 동남구 신부3길 16-1, 3층(신부동)
이현진
경기도 광주시 곤지암읍 곤지암로11번길 11-10, 108동 1401호(삼주노블리제아파트)
(74) 대리인
특허법인오암

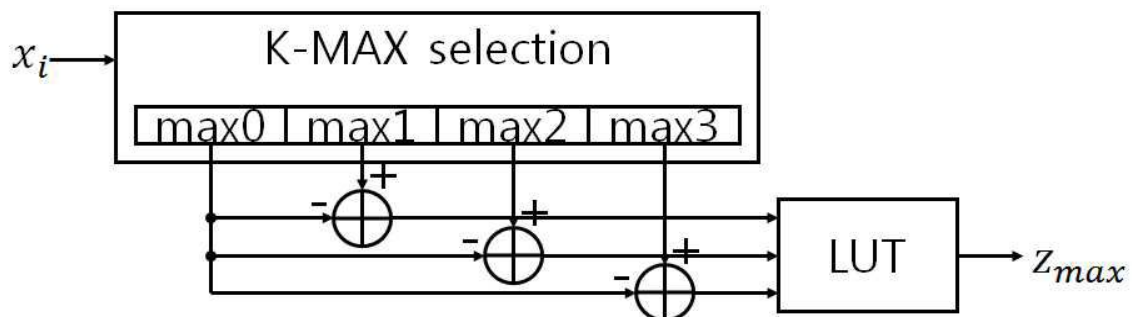
전체 청구항 수 : 총 3 항

(54) 발명의 명칭 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법 및 그 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체

(57) 요약

본 발명의 일 실시예에 따른 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법은, CNN 알고리즘 기반의 소프트맥스 계층의 연산 방법에 있어서, 소프트맥스 계층의 결과값을 도출하기 위한 N개의 입력 데이터를 입력받는 단계; 상기 입력 데이터 중 K개의 가장 큰 값들을 선택하여 LUT에 입력하는 단계; 상기 LUT에 입력된 데이터 및 소프트맥스 함수를 이용하여 상기 결과값을 산출하는 단계를 포함하되, 상기 LUT에 입력되는 데이터의 수는 선택된 K개의 가장 큰 값들에서 입력 최대값을 감산하여 LUT 입력 데이터의 수를 줄여 연산량을 감소시킬 수 있도록 하는 것을 특징으로 한다.

대 표 도 - 도3



(52) CPC특허분류

G06N 3/063 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345323366
과제번호	LINCPLUS-2020-45
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	사회맞춤형산학협력선도대학육성(LINC+)
연구과제명	자율주행 자동차를 위한 저전력 비전 인식 시스템
기 여 율	1/1
과제수행기관명	한국기술교육대학교 산학협력단
연구기간	2020.06.01 ~ 2020.12.31

명세서

청구범위

청구항 1

CNN 알고리즘 기반의 소프트맥스 계층의 연산 방법에 있어서,

소프트맥스 계층의 결과값을 도출하기 위한 N개의 입력 데이터를 입력받는 단계;

상기 입력 데이터 중 기준값에 따라 설정된 K개의 가장 큰 값들을 선택하여 LUT에 입력하는 단계;

상기 LUT에 입력된 데이터 및 소프트맥스 함수를 이용하여 상기 결과값을 산출하는 단계를 포함하되,

상기 LUT에 입력되는 데이터의 수는 선택된 K개의 가장 큰 값들에서 입력 최대값을 감산하여 LUT 입력 데이터의 수를 줄여 연산량을 감소 시킬수 있도록 하는 것을 특징으로 하는 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법.

청구항 2

제1항에 있어서,

상기 결과값($z_{\max}$)은 하기의 수학식에 의거하여 계산되는 것을 특징으로 하는 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법.

[수학식]

$$z_{\max} = \frac{e^{x_{\max}}}{\sum_{j=0}^{N-1} e^{x_j}} = \frac{1}{\sum_{j=0}^{N-1} e^{x_j - x_{\max}}} \\ \simeq \frac{1}{1 + \sum_{k=1}^{K-1} e^{x_{\max_k} - x_{\max}}}$$

위 수학식에서 입력 데이터(x)에서 $\max_k$ 는 K개의 가장 큰 값들의 첨자들이며, 입력 데이터의 최대값의 첨자는 $\max$ 이다.

청구항 3

제1항 또는 제2항에 기재된 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

발명의 설명

기술 분야

[0001] 본 발명은 CNN 알고리즘 연산회로의 연산 방법에 관한 것으로, 더욱 상세하게는 CNN 알고리즘의 소프트 맥스(Softmax) 계층의 효율적인 연산 방법에 관한 것이다.

배경 기술

[0002] 최근 다양한 영상 인식 등의 분야에서 합성곱 신경망(Convolutional Neural Network, 이하 CNN)이 높은 성능을 보이며, 널리 사용되고 있다.

[0003] 그러나 CNN이 많은 양의 연산을 요구함에 따라 임베디드 환경에서 사용되기에는 어려움이 있다.

[0004] 이러한 단점을 극복하기 위해 CNN을 위한 전용 가속 회로를 설계하는 연구가 많이 이루어지고 있으며, CNN을 구성하는 여러 종류의 연산들은 대부분 가속 회로에서 구현하기에 적합하다.

[0005] 한편 CNN은 도 1을 참조하면, 2차원 CNN의 기본적인 계층 구조를 확인할 수 있다. 구체적으로, 2차원 CNN은 입

력 이미지를 입력으로 받아 합성곱 연산을 수행하여 특징맵(feature map)을 출력하는 컨볼루션 계층(Convolutional layer, 10), 활성화 함수(activation function)를 이용하여 컨볼루션 계층(10)의 출력값을 정규화하는 활성화 계층(activation layer, 11), 활성화 계층(11)의 출력에 대하여 샘플링 또는 풀링을 수행하여 대표적 특징을 추출하는 풀링 계층(pooling layer, 12)를 포함할 수 있다. 이때, 컨볼루션 계층(10), 활성화 계층(11), 풀링 계층(12)의 연결 구조는 여러 개를 반복적으로 구성될 수 있다.

[0006] 또한, CNN은 상기 연결 구조의 후단에 풀링 계층(12)을 통해 추출된 여러 개의 특징들을 결합하는 전결합층(Fully-Connected layer, 13)이 연결되고, 소프트맥스 함수(softmax function)를 이용하여 전결합층(13)의 출력값을 정규화하는 소프트 맥스 계층(Softmax layer, 14)으로 연결될 수 있다. 전결합층(13)은 입력된 특징맵의 모든 픽셀을 벡터화하고, 각각의 파라미터 값들을 곱한 후, 연산결과들을 종합하여 가장 큰 값을 갖는 클래스를 결과로 출력할 수 있다. 소프트맥스 계층(14)은, 소프트맥스 함수를 사용하여 전결합층(13)에서의 연산 결과 값을 0과 1 사이의 확률 값으로 표현할 수 있다. 예를 들어, 소프트맥스 함수는 입력된 값을 0~1사이의 값으로 모두 정규화하며 출력 값들의 총합은 항상 1이 되는 특성을 가진 함수일 수 있다. 도 1에서, 소프트맥스 계층(14)은 설명의 편의를 위해 전결합층(13)과 별도로 도시하였으나 전결합층(13)에 포함되는 것으로 해석할 수도 있다.

[0007] 그러나 CNN에서 인식된 객체의 종류에 대해 확률을 계산하기 위해 사용되는 소프트맥스 계층의 연산의 경우, 고속 회로에서 구현하기 어려운 지수 연산과 나눗셈 연산으로 구성되어 있다.

[0008] 소프트맥스 연산을 구성한 지수 연산과 나눗셈 연산을 구현하기 위해 종래에 많이 이용한 방법은 나눗셈 연산을 로그 연산과 뺄셈으로 변환하여 구현하는 방법이다. 이 경우 소프트맥스 연산은 지수 연산과 로그 연산으로 구성되며, 각각을 LUT(Look-Up Table, 이하 LUT라 함)로 구현하여 연산한다. 그러나 이 경우 LUT의 개수가 많아지므로 이를 줄이기 위해 필요한 최소 입력 데이터의 일부를 사용하여 LUT을 줄여서 연산하는 효율적인 연산 방법이 필요하게 되었다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 한국공개특허 제10-2020-0092509호(2020년08월04일 공개)
(특허문헌 0002) (비특허문헌 001) A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proceedings of Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

발명의 내용

해결하려는 과제

[0010] 본 발명의 목적은 입력 데이터에 의해 필요한 지수 연산 및 로그 연산에 대하여 LUT의 개수를 줄여 연산량을 줄일 수 있는 CNN 알고리즘 기반의 소프트 맥스층의 효율적인 연산 방법을 제공하는 것이다.

과제의 해결 수단

[0011] 본 발명의 일 실시예에 따른 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법은, CNN 알고리즘 기반의 소프트맥스 계층의 연산 방법에 있어서, 소프트맥스 계층의 결과값을 도출하기 위한 N개의 입력 데이터를 입력 받는 단계; 상기 입력 데이터 중 기준값에 따라 설정된 K개의 가장 큰 값들을 선택하여 LUT에 입력하는 단계; 상기 LUT에 입력된 데이터 및 소프트맥스 함수를 이용하여 상기 결과값을 산출하는 단계를 포함하되, 상기 LUT에 입력되는 데이터의 수는 선택된 K개의 가장 큰 값들에서 입력 최대값을 감산하여 LUT 입력 데이터의 수를 줄여 연산량을 감소시킬수 있도록 하는 것을 특징으로 한다.

[0012] 상기 결과값($z^{\max}$)은 하기의 수학식에 의거하여 계산되는 것을 특징으로 한다.

[0013] [수학식]

$$z_{\max} = \frac{e^{x_{\max}}}{\sum_{j=0}^{N-1} e^{x_j}} = \frac{1}{\sum_{j=0}^{N-1} e^{x_j - x_{\max}}} \\ \simeq \frac{1}{1 + \sum_{k=1}^{K-1} e^{x_{\max} - x_{\max}}}$$

[0014]

[0015] 위 수학식에서 입력 데이터(x)에서 $\max_k$ 는 K개의 가장 큰 값들의 첨자들이며, 입력 데이터의 최대값의 첨자는 $\max$ 이다.

[0016] 상기 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법은 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체에 의해 수행된다.

발명의 효과

[0017] 본 발명의 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법은 CNN에 이용되는 소프트맥스 계층의 연산을 구현할 때 LUT 개수를 줄여, 연산량을 감소시킴으로써, 연산 회로의 복잡도를 줄여, 효율적인 연산이 가능한 효과가 있다.

[0018] 또한, 계산의 정확도는 조금 낮아졌으나, 회로의 복잡도는 크게 낮아진 결과를 통해 소프트맥스의 연산 회로는 빠르고, 효율적인 연산이 필요한 다양한 CNN 알고리즘에서 널리 이용될 수 있어 기대효과가 큰 장점이 있다.

도면의 간단한 설명

[0019] 도 1은 종래의 CNN 알고리즘 개념을 설명한 도면이다.

도 2는 종래 CNN 알고리즘 기반의 소프트 맥스 층의 연산 회로를 나타낸 블록도이다.

도 3은 본 발명의 CNN 알고리즘 기반의 소프트 맥스 층의 연산 회로를 나타낸 블록도이다.

도 4는 본 발명의 일 실시예에 따른 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법의 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 이하에서는 도면을 참조하여 본 발명의 구체적인 실시예를 상세하게 설명한다. 다만, 본 발명의 사상은 제시되는 실시예에 제한되지 아니하고, 본 발명의 사상을 이해하는 당업자는 동일한 사상의 범위 내에서 다른 구성요소를 추가, 변경, 삭제 등을 통하여, 퇴보적인 다른 발명이나 본 발명 사상의 범위 내에 포함되는 다른 실시예를 용이하게 제안할 수 있을 것이나, 이 또한 본원 발명 사상 범위 내에 포함된다고 할 것이다. 또한, 각 실시예의 도면에 나타나는 동일한 사상의 범위 내의 기능이 동일한 구성요소는 동일한 참조부호를 사용하여 설명한다.

[0021] 도 2는 종래 CNN 알고리즘 기반의 소프트 맥스 층의 연산 회로를 나타낸 블록도이며, 도 3은 본 발명의 CNN 알고리즘 기반의 소프트 맥스 층의 연산 회로를 나타낸 블록도이다.

[0022] 소프트맥스 계층은 일반적으로 CNN의 마지막에 사용되며, N개의 입력 값 x_i 에 대해 다음 수학식 1과 같은 연산을 하여 0과 1사이의 값을 출력한다. 모든 N개의 값에 대해 이 연산을 할 수도 있으나, CNN에서 주로 가장 큰 값만을 요구하므로 본 발명에서는 효율적인 연산을 위해 가장 큰 값만을 출력하는 경우를 가정한다.

수학식 1

$$z_i = \frac{e^{x_i}}{\sum_{j=0}^{N-1} e^{x_j}}$$

[0023]

[0024] 소프트맥스 계층에서 사용되는 지수 연산과 나눗셈은 모두 회로로 구현하기에 적합하지 않은 연산들이다. 이 연산들을 회로로 구현할 때는 수학식 2와 같이 지수 연산과 로그 연산으로 변환한 뒤, 각 연산을 LUT로 만들어 구

현할 수 있다.

수학식 2

$$\log z_i = \log \frac{e^{x_i}}{\sum_{j=0}^{N-1} e^{x_j}} = x_i - \log \sum_{j=0}^{N-1} e^{x_j}$$

[0025]

[0026] 위 수학식 2와 같이 연산한 결과에 다시 지수 연산을 취하여 마지막 결과인 z_i 를 계산한다. 이러한 과정을 LUT을 포함한 연산회로로 구현하면 도 2와 같다. 도 2에서 EXP와 LOG는 지수 연산과 로그 연산을 수행하는 LUT들이며, MAX는 최댓값을, SUM은 합을 구하는 연산 회로를 나타낸다.

[0027] 종래의 연산 회로에서는 이와 같이 LUT를 3개나 사용해야 하며, 특히 로그 연산을 구현하는 LUT는 입력이 지수 연산의 결과이므로 비트 폭이 커서 매우 많은 엔트리가 필요하다.

[0028] 이를 극복하기 위해 본 발명에서는 LUT를 한 개만 이용하는 구조를 제안한다. 입력 값들 중 K개의 제일 큰 값들을 골라서 LUT에 입력으로 넣으며, LUT는 이들로부터 결과값을 계산한 값들을 직접 가지고 있다. 여기서, 가장 큰 값에 해당하는 K개를 결정하는 기준은 설정된 기준값에 의해 정해질 수 있으며, 기준값은 CNN 알고리즘 적용 분야나 연산량 등에 따라 달라질 수 있다.

[0029] 특히 K개의 제일 큰 값들을 선택한 뒤 LUT에 넣기 전에 최댓값에서 빼 주면, 아래와 같은 수학식 3에 의거하여 LUT의 입력으로 사용되는 값의 개수를 1개 줄일 수 있다. 또한 LUT의 엔트리 수를 줄이기 위해 각 입력값들에서 상위 w 비트만을 취하여 사용한다.

수학식 3

$$z_{\max} = \frac{e^{x_{\max}}}{\sum_{j=0}^{N-1} e^{x_j}} = \frac{1}{\sum_{j=0}^{N-1} e^{x_j - x_{\max}}} \simeq \frac{1}{1 + \sum_{k=1}^{K-1} e^{x_{\max k} - x_{\max}}}$$

[0030]

[0031] 위 수학식 3에서 입력 데이터(x)에서 $\max_k$ 는 K개 가장 큰 값들의 첨자들이며, 입력 데이터의 최댓값의 첨자는 $\max$ 이다. 위 수학식 3의 값들을 LUT에 넣어서 연산회로를 구현하면 도 3과 같다.

[0032] 본 발명의 방법을 기존 방법과 두 가지 측면에서 비교하였다. 계산의 정확성을 비교하기 위해 부동 소수점으로 계산한 출력값과의 RMS(root-mean-square) 에러를 측정하여 비교하였다. 입력으로는 VOC 데이터셋을 이용한 객체 검출 상황을 가정하여 21개의 값을 임의로 생성하여 넣었으며, 1개의 값은 큰 값, 나머지는 작은 값이 되도록 생성하였다. 기존의 방식에서는 RMS 에러가 0.0271이었으며, 제안한 방법에서는 0.0646이었다. 제안한 방법에서는 4개의 큰 값들을 선택하였으며, 상위 3개 비트를 이용하여 LUT를 접근하였다. 기존의 방법에 비해 RMS 에러가 증가하였으나, 그 차이는 크지 않았다.

[0033] 본 발명에서 제안하는 방법을 이용할 경우의 회로 복잡도를 간접적으로 비교하기 위해 LUT의 크기를 비교하였다. 입력 비트를 8비트, 지수 연산의 결과 비트 18비트, 로그 연산의 결과를 8비트라고 가정하였을 때, 기존 방법의 LUT의 크기는 $2 \times 28 \times 18 + 218 \times 8 = 2,160,368$ 비트이다. 이에 반해 본 발명에서 제안하는 방법에서의 LUT의 크기는 $29 \times 18 = 9216$ 비트여서, 기존에 비해 LUT의 크기가 크게 줄어들었음을 알 수 있다.

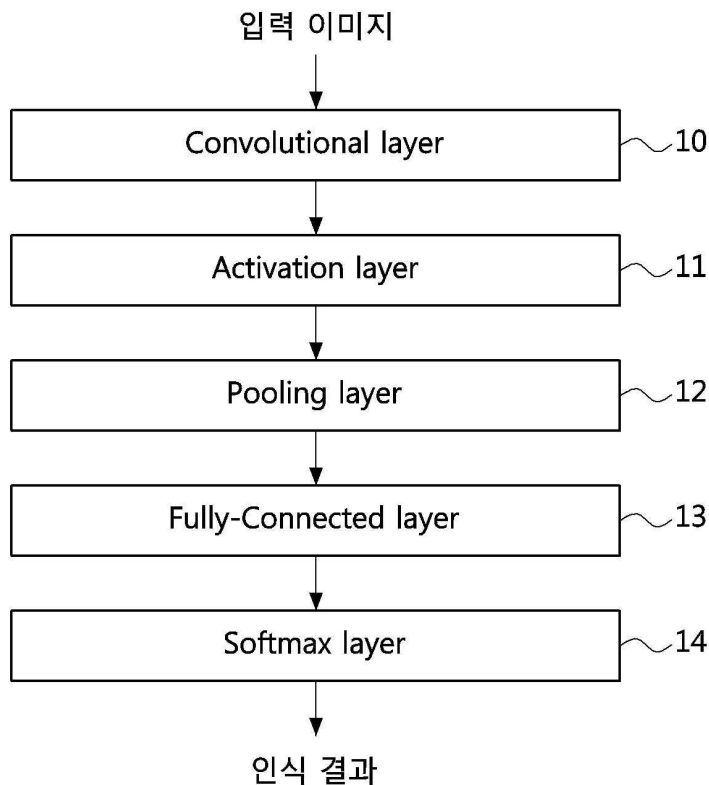
[0034] 도 4는 본 발명의 일 실시예에 따른 CNN 알고리즘의 소프트 맥스 계층의 효율적인 연산 방법의 순서도이다.

[0035] CNN 알고리즘의 전결합층은 입력된 특징맵의 모든 픽셀을 벡터화하고, 각각의 파라미터 값들을 곱한 후, 연산 결과들을 종합하여 가장 큰 값을 갖는 클래스를 결과로 출력할 수 있으며, 소프트맥스 계층은, 소프트맥스 함수를 사용하여 전결합층에서의 연산 결과 값을 0과 1 사이의 확률 값으로 표현한다.

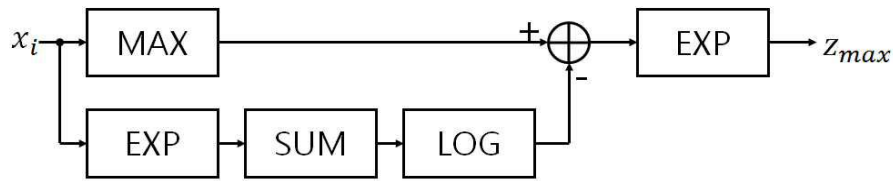
- [0036] 소프트맥스 계층의 효율적인 연산을 위해 본 발명은, 우선 소프트 계층은 N개의 입력 데이터를 입력받는다(S10).
- [0037] 입력 데이터 중 K개의 가장 큰 값들을 선택하여 LUT에 입력시킨다(S12).
- [0038] 이때, LUT에 입력되는 데이터의 수는 수학적 식 3에 표현된 바와 같이, 선택된 K개의 가장 큰 값들에서 입력 최대 값을 감산($x_{\max_k} - x_{\max}$)하여 LUT 입력 데이터의 수를 줄일 수 있게 된다.
- [0039] LUT는 입력된 데이터와 소프트 맥스 함수를 이용하여 결과값 $z_{\max}$ 을 산출한다(S14).
- [0040] 상기 결과값 $z_{\max}$ 은 상술한 수학적 식 3에 의거하여 계산된다.
- [0041] 나아가 도 4의 결과값 $z_{\max}$ 을 도출하기 위해 연산하는 일련의 과정은 컴퓨터에서 처리 가능한 컴퓨터 프로그램이 내장된 프로그램수행모듈에 의해 수행될 수 있다.
- [0042] 또한, 프로그램수행모듈은 컴퓨터 프로그램 자체가 될 수 있으며, 컴퓨터가 판독 가능한 기록 매체에 저장되어 수행될 수도 있다.
- [0043] 또한 프로그램수행모듈은 중앙 관리 컴퓨터 또는 현장 컴퓨터에 구비될 수 있으며, 프로그램수행모듈은 컴퓨터 프로그램이 내장된 하드웨어 또는 컴퓨터 프로그램이 될 수 있다.

도면

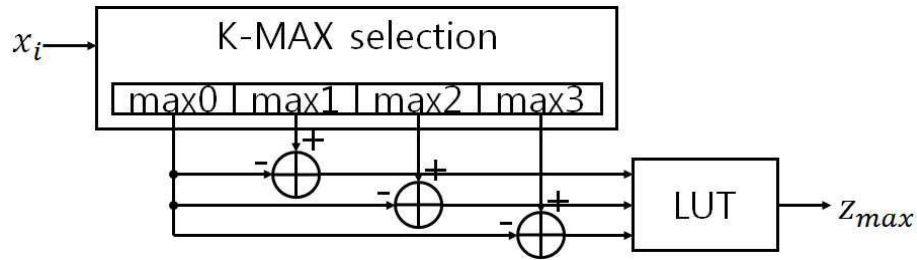
도면1



도면2



도면3



도면4

