(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0095163
(43) 공개일자 2020년08월10일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
(21) 출원번호 10-2019-0012914
(22) 출원일자 2019년01월31일
심사청구일자 2019년01월31일

(71) 출원인
한국기술교육대학교 산학협력단
충청남도 천안시 동남구 병천면 충절로 1600 (한국기술교육대학교내)
(72) 발명자
강형주
대전광역시 서구 둔산북로 160, 4동 1204호 (둔산동, 한마루아파트)
우용근
전라남도 목포시 하당로 44번길 5-2
(74) 대리인
특허법인오암

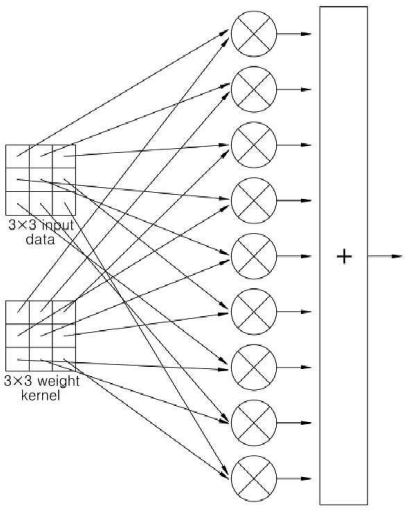
전체 청구항 수 : 총 3 항

(54) 발명의 명칭 **가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치**

(57) 요 약

가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치가 개시된다. 합성곱 신경망의 연산 회로에 있어서, 입력 데이터(10)에서 X 배열 또는 + 배열을 선택하는 먹스(100); 먹스(100)의 출력과 신경망의 가중치를 곱하는 곱셈기(200); 및 곱셈기(200)의 출력을 더하는 덧셈기(300)를 포함한다. 따라서 곱셈기(200) 수를 줄여 고속 회로를 만들 수 있고, 연산 회로에서 가속 회로의 면적이 줄어 드는 장점이 있다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

합성곱 신경망의 연산 회로에 있어서,

입력 데이터(10)에서 X 배열 또는 + 배열을 선택하는 먹스(100);

상기 먹스(100)의 출력과 신경망의 가중치를 곱하는 곱셈기(200); 및

상기 곱셈기(200)의 출력을 더하는 덧셈기(300)를 포함하는 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치.

청구항 2

제1항에 있어서,

상기 입력 데이터(10)는 3*3 배열이고, 상기 X배열은 대각선과 교차점의 입력 데이터(10)이고, 상기 + 배열은 십자선과 십자점의 입력 데이터(10)이고, 상기 먹스(100)는 2입력1출력 먹스(100)가 4개이고, 상기 곱셈기(200)와 상기 가중치의 수는 5개인 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치.

청구항 3

제1항에 있어서,

상기 Conv-XP 프루닝 장치를 사용하여 연산 회로인 가속 회로를 구성하는 합성곱 신경망.

발명의 설명

기 술 분 야

[0001] 본 발명은 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치에 관한 것으로, 더욱 상세하게는 곱셈기 수를 줄이는 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치에 관한 것이다.

배 경 기 술

[0002] 합성곱 신경망은 입력 데이터와 가중치가 곱셈기에 의해 곱해지고, 곱셈 결과가 덧셈기로 더해져 출력한다. 합성곱 신경망의 연산 회로는 입력 데이터와 가중치를 곱하는 곱셈기와 곱셈 결과를 더하는 덧셈기로 구성된다. 이러한 연산 회로를 예를 들어 살펴 보면 다음과 같다.

[0003] 도 1은 종래 합성곱 신경망의 연산 회로를 보인 예시도이다.

[0004] 종래 합성곱 신경망의 연산 회로는 3*3 입력 데이터와 3*3 가중치를 9개의 곱셈기가 곱한 후 덧셈기가 더한다. 이러한 연산 회로는 9개의 곱셈기를 가지고 있고, 곱셈기는 복잡도가 높아 전체 연산 회로의 면적이 증가하는 문제점이 있다.

[0005] 도 2는 종래 합성곱 신경망의 다른 실시예를 보인 예시도이다.

[0006] 종래 합성곱 신경망의 연산 회로는 3*3 입력 데이터를 다중화하는 5개의 먹스와 5개의 먹스 출력과 5개의 가중치를 곱하는 5개의 곱셈기를 가지고, 하나의 덧셈기를 가진다. 이러한 연산 회로는 5개의 곱셈기를 가지지만 먹스 또한 5개를 가져 연산 회로의 복잡도가 증가하는 문제점이 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 공개번호 제10-2018-0123846호, 합성곱 신경망을 위한 논리적 3차원 구조의 재구성형 연산 가

속기

발명의 내용

해결하려는 과제

[0008] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은, 곱셈기 수를 줄이는 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치를 제공하는데 있다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명은, 합성곱 신경망의 연산 회로에 있어서, 입력 데이터(10)에서 X 배열 또는 + 배열을 선택하는 먹스(100); 먹스(100)의 출력과 신경망의 가중치를 곱하는 곱셈기(200); 및 곱셈기(200)의 출력을 더하는 덧셈기(300)를 포함한다.

[0010] 또한, 입력 데이터(10)는 3*3 배열이고, X배열은 대각선과 교차점의 입력 데이터(10)이고, + 배열은 십자선과 십자점의 입력 데이터(10)이고, 먹스(100)는 2입력1출력 먹스(100)가 4개이고, 곱셈기(200)와 가중치의 수는 5개이다.

[0011] 또한, 먹스(100)가 입력 데이터(10)에서 대각선 입력 데이터(10) 또는 십자선 입력 데이터(10)의 값에 따라 입력 데이터(10)에 대한 먹스(100) 출력을 전환한다. 먹스(100)가 대각선 입력 데이터(10)와 십자선 입력 데이터(10)의 합값이 가지는 레벨에 따라 먹스(100) 출력을 전환할 수 있다. 또한, 입력 데이터(10)에 따른 합성곱 신경망 출력이 피드백되어 가중치가 학습될 수 있다. 이때, 가중치가 입력 데이터(10)의 X 배열 또는 + 배열 선택에 따라 달라질 수 있다.

[0012] 또한, Conv-XP 프루닝 장치를 사용하여 연산 회로인 가속 회로를 구성하는 합성곱 신경망이다.

발명의 효과

[0013] 상기와 같은 본 발명에 따른 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치를 이용할 경우에는 곱셈기(200) 수를 줄여 고속 회로를 만들 수 있다.

[0014] 또한, 연산 회로에서 가속 회로의 면적이 줄어 드는 장점이 있다.

도면의 간단한 설명

[0015] 도 1은 종래 합성곱 신경망의 연산 회로를 보인 예시도이다.

도 2는 종래 합성곱 신경망의 다른 실시예를 보인 예시도이다.

도 3은 X 배열과 + 배열을 설명한 예시도이다.

도 4는 본 발명 합성곱 신경망의 연산 회로를 보인 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0016] 명세서에서 사용된 용어는 그 단어가 본래 가지는 의미가 확대 해석될 수 있으며 서로 다른 의미를 가지는 단어와 단어가 하나의 단어로 조합됨으로써 광의의 새로운 의미를 가지는 합성어 형태를 취한다. 예를 들어 합성곱이라는 단어와 신경망이라는 단어가 합성되어 합성곱 신경망을 의미하게 된다. 또한, 명세서에서 제시하는 실시예는 실시에 바람직한 실시예이며, 경우에 따라 다른 구성이 부가되거나 본래 있던 구성이 생략됨이 가능하다. 실시예에서 구성을 포함하다 또는 가지다로 실시될 수 있다.

[0017] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0019] 도 3은 X 배열과 + 배열을 설명한 예시도이다.

[0020] 3*3 입력 데이터(10)에서 X 배열은 대각선으로 0이 아닌 입력 데이터(10)를 가지고, + 배열은 십자선으로 0이 아닌 입력 데이터(10)를 가진다. 3*3 입력 데이터(10)에서 대각선 또는 십자선으로 0이 아닌 입력 데이터(10)를

가지는 X 배열 또는 + 배열이 구성된다. X 배열 또는 + 배열로 특정되는 3*3 입력 데이터(10)는 합성곱 신경망의 연산 회로에 입력된다. 합성곱 신경망의 연산 회로는 X 배열 또는 + 배열에 적응해서 곱셈과 덧셈 연산을 수행한다.

[0021] 3*3 입력 데이터(10)에서 X 배열 또는 + 배열로 만드는 것이 프루닝 기법이다. 프루닝 기법은 특정 배열만 0이 아닌 값으로 만들고 0인 값은 연산에서 제외해서 합성곱 신경망의 연산 회로를 가속하는데 필수적이다. 프루닝 기법이 사용되어 합성곱 신경망을 고속 회로로 만들 수 있다.

[0022] Conv-XP 프루닝 장치는 도 1과 도 2에서 동시에 처리하는 단위인 한 커널 내의 계수 9개 중에서 일부 계수를 프루닝한다. 종래 프루닝 기법에서는 9개의 계수 중에서 몇 개의 계수가 프루닝될 지, 어떤 계수들이 0이 될 지가 정해져 있지 않다. 이에 반해 Conv-XP 프루닝 장치에서는 도 3과 같이 두 가지 패턴으로만 프루닝을 한다. 도 3에서 흰색 칸은 0으로 프루닝되는 계수를, 회색은 프루닝 후에 0이 아닌 값으로 남는 계수를 의미한다. 두 개의 패턴은 각각 글자 X와 기호 +를 닮아서 Conv-XP 프루닝이라고 부른다.

[0023] 종래 프루닝 기법에 비해 Conv-XP 프루닝 장치는 합성곱 신경망을 가속 회로에 적합하도록 프루닝한다. Conv-XP 프루닝 장치에서는 가속 회로에서 동시에 처리되는 각각의 3*3 커널에 대해 항상 4개의 계수가 0이 되도록 프루닝한다. 그러므로 5개의 곱셈기로도 항상 3*3 커널을 한 싸이클 안에 처리할 수 있음이 보장된다. 그리고 X 배열과 + 배열만 사용하므로 한 곱셈기에는 두 개의 입력 데이터만이 피연산자가 될 가능성이 있다. 이러한 특성을 이용한 합성곱 신경망의 연산 회로는 도 4와 같다.

[0024] Conv-XP 프루닝 장치는 각각의 3*3 커널에 대해 적용되고, 각각의 3*3 커널에 대해 X 배열과 + 배열에 있는 계수들의 절대값의 합을 구한 뒤, 그 합이 더 큰 배열을 취하고 배열에 따라 계수들을 0으로 만든다. 모든 3*3 커널에 대해 프루닝을 수행한 뒤, 0인 계수들을 0으로 유지하면서 다시 학습한다.

[0025] 도 4는 본 발명 합성곱 신경망의 연산 회로를 보인 예시도이다.

[0026] 합성곱 신경망의 연산 회로에 있어서, 가속 회로에 적합한 합성곱 신경망의 Conv-XP 프루닝 장치는 입력 데이터(10)에서 X 배열 또는 + 배열을 선택하는 먹스(100); 먹스(100)의 출력과 신경망의 가중치를 곱하는 곱셈기(200); 및 곱셈기(200)의 출력을 더하는 덧셈기(300)를 포함한다.

[0027] 입력 데이터(10)는 3*3 배열이고, X배열은 대각선과 교차점의 입력 데이터(10)이고, + 배열은 십자선과 십자점의 입력 데이터(10)이고, 먹스(100)는 2입력1출력 먹스(100)가 4개이고, 곱셈기(200)와 가중치의 수는 5개이다.

[0028] 먹스(100)가 입력 데이터(10)에서 대각선 입력 데이터(10) 또는 십자선 입력 데이터(10)의 값에 따라 입력 데이터(10)에 대한 먹스(100) 출력을 전환한다. 먹스(100)가 대각선 입력 데이터(10)와 십자선 입력 데이터(10)의 합값이 가지는 레벨에 따라 먹스(100) 출력을 전환할 수 있다. 또한, 입력 데이터(10)에 따른 합성곱 신경망 출력이 피드백되어 가중치가 학습될 수 있다. 이때, 가중치가 입력 데이터(10)의 X 배열 또는 + 배열 선택에 따라 달라질 수 있다.

[0029] Conv-XP 프루닝 장치를 사용하여 연산 회로인 가속 회로를 구성하는 합성곱 신경망이 만들어진다.

[0030] Conv-XP프루닝 장치를 사용했을 때 또 하나의 이점은 메모리 면적을 줄일 수 있다는 것이다. 종래 프루닝 기법에서는 각각의 0이 아닌 계수들이 곱해질 입력 데이터를 선택하기 위해, 각 곱셈기 앞에 위치하는 9-to-1 MUX의 선택신호를 같이 저장해야 한다. 한 개의 MUX에 4bit의 선택신호가 필요하고, 5개의 곱셈기를 가정하면 총 20bit의 선택신호가 필요하다.

[0031] 그러나 Conv-XP프루닝 장치를 사용하게 되면 X 배열과 + 배열 사이에만 선택하므로 각 곱셈기 앞에는 2-to-1 먹스만 있으며 되고 이 먹스의 선택 신호 길이는 1bit로 충분하다. 그리고 5개의 곱셈기 앞에 있는 먹스들이 같은 선택신호를 공유하므로 5개의 0이 아닌 계수가 하나의 선택 신호를 공유하게 된다. 그러므로 5개의 곱셈기에 대해 20bit의 선택신호가 1bit로 줄어들게 되고, 따라서 이 선택 신호들을 저장하는 메모리의 크기가 줄어들게 된다.

표 1

[0032]

	VGG16	ResNet-50
baseline	88.44%	91.14%
Conv-XP	89.74%	91.30%

[0033] 표1에서 VGG16과 ResNet-50에 대해 Conv-XP프루닝 장치를 적용하기 전의 정확도('baseline')와 Conv-XP프루닝 장치를 적용한 후 재학습했을 때의 정확도('Conv-XP')를 비교한다. 재학습은 학습 속도를 0.001로 하고 12 epoch 동안 학습하였다. 표1의 결과에 따르면 Conv-XP프루닝 장치를 적용하더라도 합성곱 신경망의 성능이 저하되지 않음을 알 수 있다. Conv-XP프루닝 장치를 사용했을 때 가속 회로의 복잡도 개선을 비교하기 위해서 세 가지의 회로를 RTL(register transfer level)에서 설계한 뒤 합성하여 비교하였다. 첫 번째는 프루닝 되지 않은 합성곱 신경망을 가정한 도 1과 같은 기존의 구조이고, 두 번째는 기존의 프루닝 기법이 적용된 합성곱 신경망을 위한 sparse 가속 회로인 도 2와 같은 구조이며, 세 번째는 Conv-XP프루닝 장치로 프루닝된 합성곱 신경망을 위한 sparse 가속 회로인 도 4와 같은 구조이다. 표 2에서는 각각을 'No sparsity', 'Conventional sparsity', 'Conv-XP'로 표기하였다. 합성 라이브러리는 Global Foundry사의 65nm 공정을 이용하였다. 표 2에서 PE(processing element)는 9개 또는 5개의 곱셈기(200)로 이루어져서 하나의 3x3 커널을 동시에 처리할 수 있는 블록을 의미하며, 이러한 PE가 16개 있는 것을 가정하였다.

표 2

[0034]

	16 PEs	Weight Buffer	Index Buffer	Activation MUX	Miscellaneous	Total
No sparsity	642826	507042	-	-	3199	1153067
Conventional sparsity	357640	281691	81895	25078	56	746360
Conv-XP	357364	281691	5916	6157	47	651175

[0035] 기존의 프루닝 기법을 위한 sparse 가속 회로(표2에서 'Conventional sparsity')는 프루닝을 고려하지 않는 일반적인 가속 회로(표 2에서 'No sparsity')에 비해 곱셈기(200)의 개수가 9개에서 5개로 줄고 필요한 계수의 개수도 9개에서 5개로 감소하므로, PE와 계수 버퍼(Weight Buffer)의 면적이 감소하였다. 그러나 그림 2에 그려져 있는 9-to-1 MUX들을 모아 놓은 Activation MUX가 추가되고, MUX 선택신호를 저장하는 Index Buffer도 역시 추가된다.합성곱 신경망을 Conv-XP프루닝 장치로 프루닝하면 sparse 가속 회로에서 Activation MUX와 Index Buffer의 크기를 줄일 수 있다. 기존의 프루닝 기법에서는 프루닝의 패턴이 없으므로 9개의 입력 데이터(10) 모두로부터 1개의 데이터를 선택하는 9-to-1 MUX가 필요했지만, Conv-XP에서는 두 가지의 패턴 중 하나로 프루닝하므로 2-to-1 MUX로 충분하다. 그러므로 Activation MUX 블록의 크기가 약 25%로 감소된다. 그리고 MUX 선택 신호의 비트 폭도 4 비트에서 1 비트로 줄어들고, 5개의 계수들이 하나의 선택신호를 공유하므로 결국 20 비트에서 1 비트로 줄어든다. 그 결과 Index Buffer의 크기가 약 7.2%로 감소하였다. 이러한 효과들로 인해 Conv-XP프루닝 장치를 가정하면 sparse 가속 회로의 면적이 기존의 sparse 가속 회로에 비해 약 12.8% 감소한다.

[0037] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

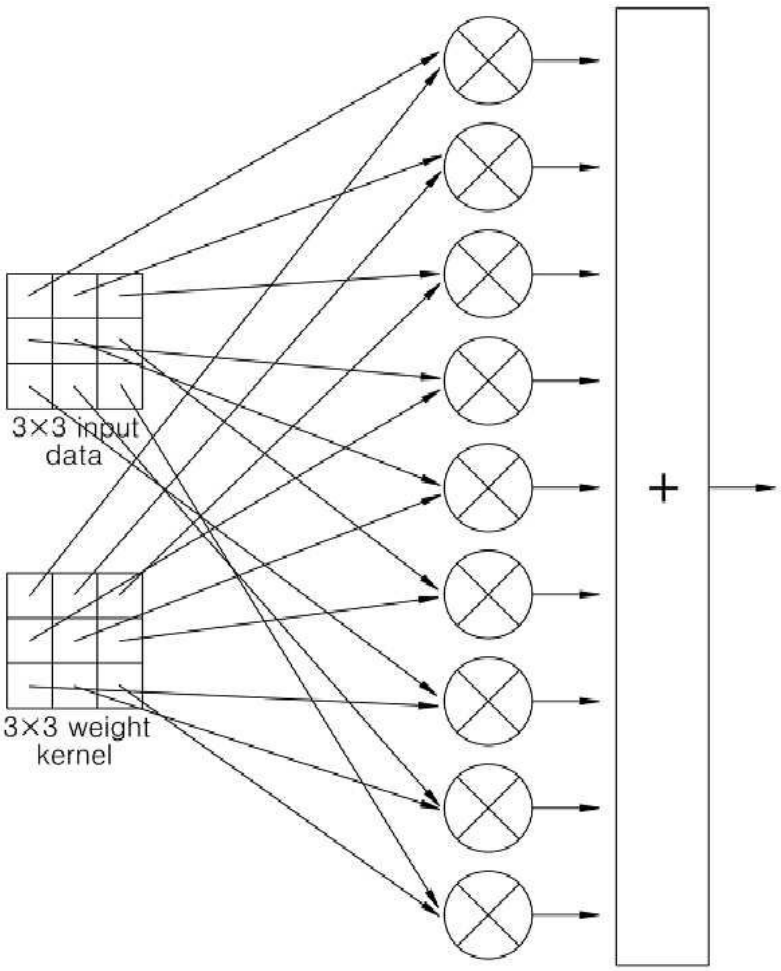
부호의 설명

[0038] 100: 먹스 200: 곱셈기

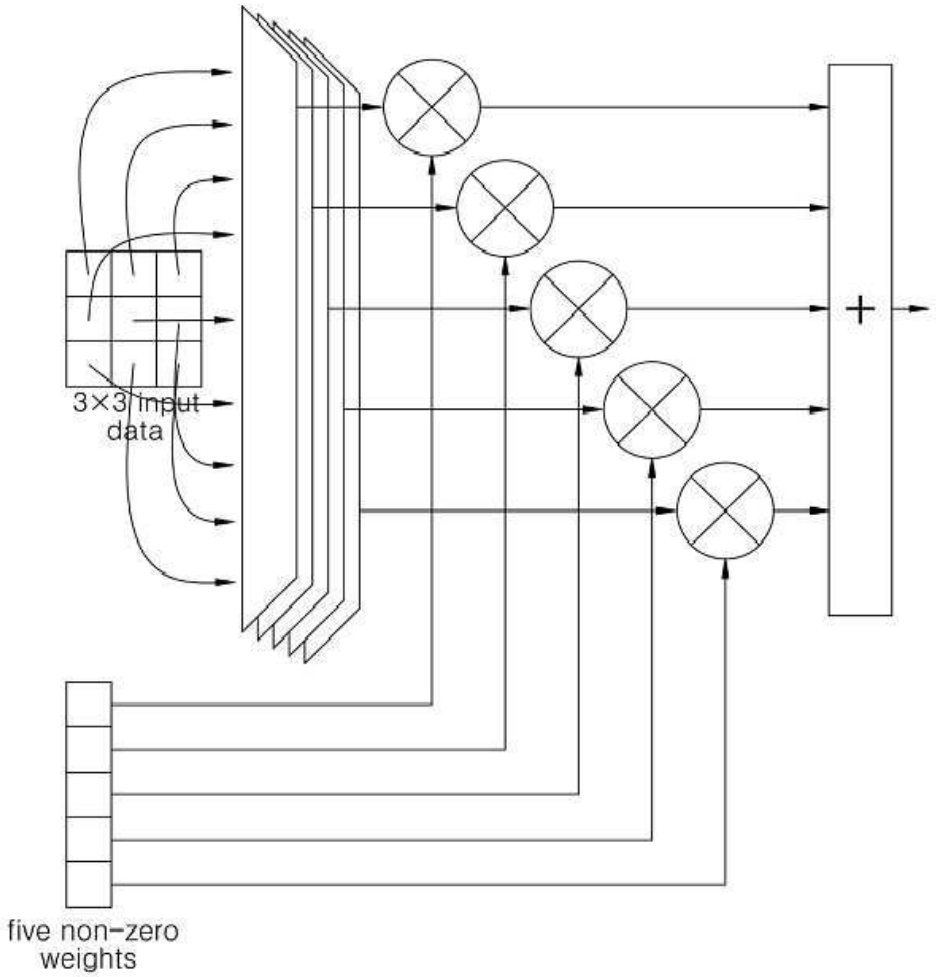
300: 덧셈기

도면

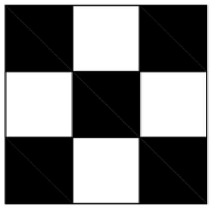
도면1



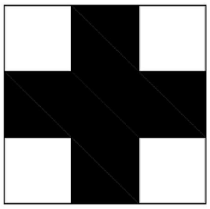
도면2



도면3

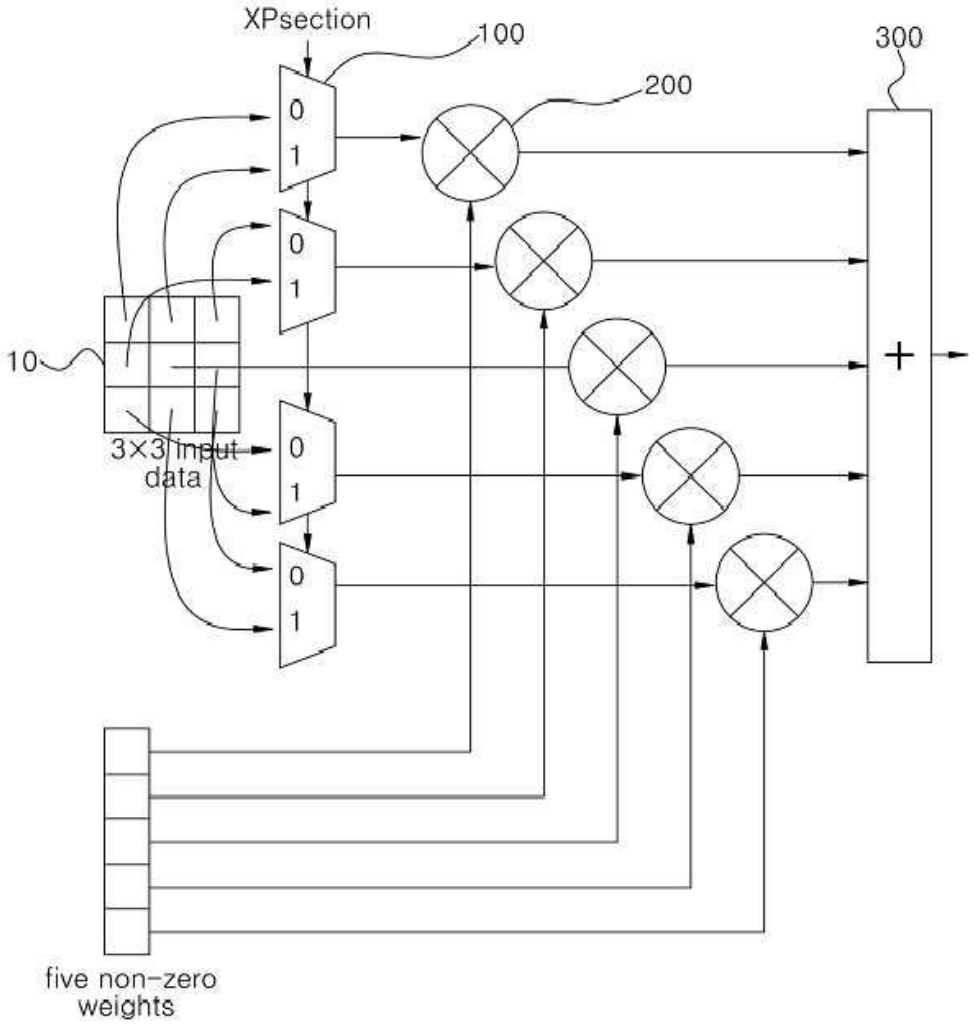


(a)



(b)

도면4


XPsection
100
200
300
10
3×3 input data
0
1
five non-zero weights