



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0038414
(43) 공개일자 2020년04월13일

(51) 국제특허분류(Int. Cl.)
G06F 7/544 (2017.01) G06N 3/04 (2006.01)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06F 7/5443 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2019-0118500
(22) 출원일자 2019년09월26일
심사청구일자 없음
(30) 우선권주장
JP-P-2018-188612 2018년10월03일 일본(JP)

(71) 출원인
캐논 가부시끼가이샤
일본 도쿄도 오오따꾸 시모마루코 3조메 30방 2고
(72) 발명자
첸 체웨이
일본국 도쿄도 오오따꾸 시모마루코 3조메 30방
2고 캐논 가부시끼가이샤 나이
(74) 대리인
권태복

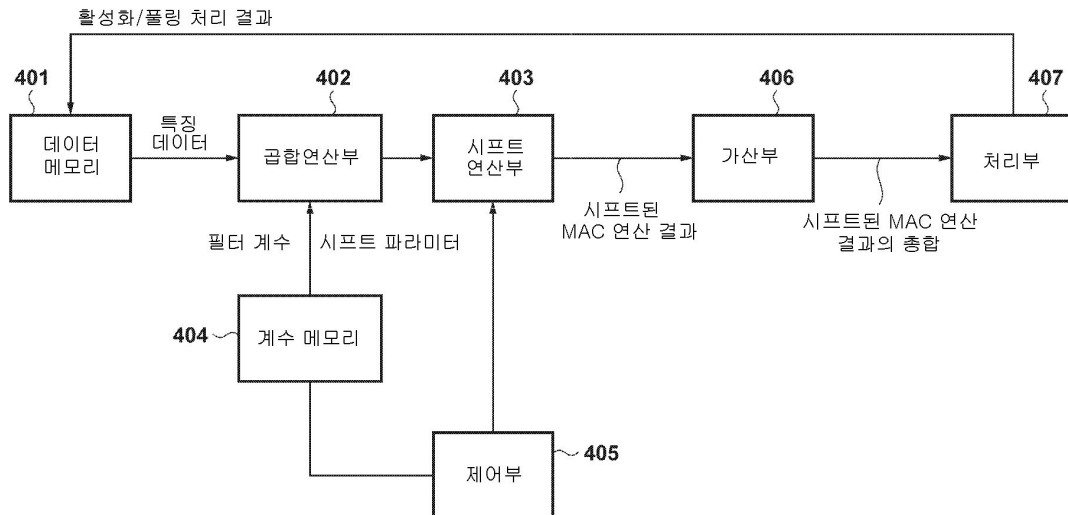
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 정보 처리장치, 정보 처리방법, 비일시적인 컴퓨터 판독가능한 기억매체

(57) 요약

정보 처리장치는, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와, 상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC(multiply-accumulate) 연산을 실행하도록 구성된 복수의 MAC부와, 상기 복수의 MAC부에 의해 얻어진 복수의 MAC 연산 결과를 상기 시프트량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와, 상기 복수의 시프트 연산부에 의해 시프트된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부를 구비한다.

대표도



(52) CPC특허분류
G06N 3/063 (2013.01)

명세서

청구범위

청구항 1

복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와,

상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC(multiply-accumulate) 연산을 실행하도록 구성된 복수의 MAC부와,

상기 복수의 MAC부에 의해 얻어진 복수의 MAC 연산 결과를 상기 시프트량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와,

상기 복수의 시프트 연산부에 의해 시프트된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부를 구비한 정보 처리장치.

청구항 2

제 1항에 있어서,

상기 복수의 MAC부는 상기 복수의 계층 각각의 MAC 연산을 동일한 회로 상에서 실행하는 정보 처리장치.

청구항 3

제 2항에 있어서,

상기 복수의 MAC부는, 상기 복수의 계층 각각의 MAC 연산을 동일한 회로 상에서 순차적으로 실행하는 정보 처리장치.

청구항 4

제 3항에 있어서,

상기 제어부는 처리할 계층마다 상기 시프트량을 설정하는 정보 처리장치.

청구항 5

제 1항에 있어서,

상기 복수의 계층은, 서로 다른 비트 폭을 갖는 데이터 세트들을 처리하는 2개 이상의 계층을 포함하는 정보 처리장치.

청구항 6

제 1항에 있어서,

상기 제어부는, 상기 계층의 데이터의 비트 폭에 따라 필터 계수의 전송 횟수를 전환하는 정보 처리장치.

청구항 7

제 1항에 있어서,

상기 시프트 연산부는 복수의 시프트 량에 근거하여 상기 복수의 MAC 연산 결과를 시프트하는 정보 처리장치.

청구항 8

복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트 량을 설정하도록 구성된 제어부와,

상기 계층의 복수의 데이터를 상기 시프트 량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와,

복수의 필터 계수와, 상기 복수의 시프트 연산부에 의해 시프트된 복수의 데이터에 대해, MAC 연산을 실행하도록 구성된 복수의 MAC부와,

상기 복수의 MAC부에 의해 계산된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부를 구비한 정보 처리장치.

청구항 9

제 1항에 있어서,

상기 복수의 MAC부와 상기 복수의 시프트 연산부는 병렬로 동작하는 정보 처리장치.

청구항 10

제 1항에 있어서,

상기 가산부는, 소정의 계층의 상기 총합을 상기 소정의 계층의 다음의 계층의 데이터로서 메모리에 격납하는 정보 처리장치.

청구항 11

제 1항에 있어서,

상기 총합에 대해 활성화 처리를 행하도록 구성된 활성화부를 더 구비한 정보 처리장치.

청구항 12

제 11항에 있어서,

상기 활성화 처리의 결과에 대해 풀링 처리를 행하도록 구성된 풀링부를 더 구비한 정보 처리장치.

청구항 13

제 12항에 있어서,

상기 풀링부는, 소정의 계층에 대해 상기 풀링 처리를 행하여 얻어진 결과를 상기 소정의 계층의 다음의 계층의 데이터로서 메모리에 격납하는 정보 처리장치.

청구항 14

제 10항에 있어서,

상기 메모리에 격납된 데이터에 근거하여 동화상의 각 프레임에 대해 화상처리 및 화상 인식 중 적어도 한가지를 행하도록 구성된 유닛을 더 구비한 정보 처리장치.

청구항 15

제 1항에 있어서,

상기 네트워크는 계층마다 다른 데이터의 비트 폭을 갖는 정보 처리장치.

청구항 16

제 1항에 있어서,

상기 데이터의 비트 폭은 2비트, 4비트 또는 8비트인 정보 처리장치.

청구항 17

복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하는 단계와,

상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC 연산을 실행하는 단계와,

복수의 상기 MAC 연산의 결과를 상기 시프트량에 근거하여 시프트하는 단계와,

시프트된 상기 복수의 MAC 연산 결과의 총합을 계산하는 단계를 포함하는 정보 처리방법.

청구항 18

복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하는 단계와,

상기 계층의 복수의 데이터를 상기 시프트량에 근거하여 시프트하는 단계와,

복수의 필터 계수와 시프트된 상기 복수의 데이터에 대해, MAC 연산을 실행하는 단계와,

계산된 복수의 MAC 연산 결과의 총합을 계산하는 단계를 포함하는 정보 처리방법.

청구항 19

컴퓨터를,

복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와,

상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC 연산을 실행하도록 구성된 복수의 MAC부와,

상기 복수의 MAC부에 의해 얻어진 복수의 MAC 연산 결과를 상기 시프트량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와,

상기 복수의 시프트 연산부에 의해 시프트된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부로서 기능시키는 컴퓨터 프로그램을 기억하는 비일시적인 컴퓨터 판독가능한 기억매체.

청구항 20

컴퓨터를,

복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와,

상기 계층의 복수의 데이터를 상기 시프트량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와,

복수의 필터 계수와, 상기 복수의 시프트 연산부에 의해 시프트된 복수의 데이터에 대해, MAC 연산을 실행하도록 구성된 복수의 MAC부와,

상기 복수의 MAC부에 의해 계산된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부로서 기능시키는 컴퓨터 프로그램을 기억하는 비일시적인 컴퓨터 판독가능한 기억매체.

발명의 설명

기술 분야

[0001] 본 발명은, 복수의 계층을 갖는 네트워크에 있어서의 연산 기술에 관한 것이다.

배경 기술

[0002] 최근, 심층학습의 진보에 의해, 화상 인식의 정밀도가 상승하고 있다. 컨볼루션 뉴럴 네트워크(convolutional neural networks: CNN)는, 심층학습에 사용되고 있는 수법으로서 알려져 있다. CNN에서는, 복수의 레이어가 계층적으로 접속되어 있고, 각 레이어 내부에 복수개의 특징 화상이 포함된다. 도2에, 레이어의 수가 4(레이어 1 내지 4)이며, 각 레이어 내부에 특징 화상이 4매 있는 네트워크의 예를 나타낸다. 도2에 있어서, 특징 화상(i, j)은, 레이어 i에 있어서의 j매째의 특징 화상을 나타낸다. 학습한 필터 계수와 특징 화상의 화소(특징 데이터)를 사용해서 필터 처리의 결과를 계산한다. 필터 처리는 곱합(multiply-accumulate: MAC) 연산이며, 복수의 승산과 누적 가산을 포함하고 있다. 도2의 각각의 화살표는 MAC 연산을 나타낸다.

[0003] 현재 레이어의 특징 화상은, 이전 레이어의 특징 화상과 이전 레이어에 대응하는 필터 계수를 사용해서 계산한다. 현재 레이어의 1매의 특징 화상을 계산하기 위해서는, 이전 레이어의 복수개의 특징 화상의 정보가 필요하다. 현재 레이어의 각각의 특징 화상을 계산하기 위한 컨볼루션 연산의 계산식은 이하와 같다.

수학식 1

$$O_{i,j}(n) = \sum_{m=1}^M \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} (I_{i+x,j+y}(m) \times C_{x,y}(m,n))$$

[0004]

[0005] 여기에서, $O_{i,j}(n)$ 은, 현재 레이어에 있어서의 n매째의 특징 화상 중의 위치(i, j)에 대응하는 MAC 연산 결과를 나타내는 변수다. 이 수학식 1에서는, 이전 레이어 중에 특징 화상이 M매 존재하고, m매째의 특징 화상에 있어서 위치(i, j)에 있어서 특징 데이터를 $I_{i,j}(m)$ 으로 표시하고 있다. $X \times Y$ 개의 필터 계수 $C_{x,y}(m,n)$ 내지 $C_{X,Y}(m,n)$ 이 존재하고, 필터 계수는 특징 화상마다 다르다. 현재 레이어에 있어서 n매째의 특징 화상을 산출하기 위한 MAC 연산은 $(M \times X \times Y)$ 회 행해진다. 컨볼루션 연산을 행한 후에, MAC 연산 결과 $O_{i,j}(n)$ 을 사용해서 활성화 처리나 풀링(pooling) 등의 처리를 행하여, 현재 레이어의 특징 화상을 계산한다.

[0006] CNN은 다수의 MAC 연산을 필요로 하기 때문에, 휴대 단말, 차량 탑재 기기 등의 임베디드 시스템에 CNN을 적용할 경우, 효율적인 데이터 병렬 처리장치가 요구된다. 처리 데이터의 비트 폭을 삭감하면, 컨볼루션 연산 결과를 계산하는 연산부의 코스트가 줄어들기 때문에, 연산부의 병렬도(degree of parallelism: DOP)를 상승시키는 것이 가능하게 된다. Y. Li, et al., A7.663-TOPS 8.2-W Energy-efficient FPGA Accelerator for Binary Convolutional Neural Networks, Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Pages 290-291, Feb. 2017에서는, 레이어마다 데이터 비트 폭이 다른 네트워크를

처리하는 하드웨어의 구성이 제안되어 있다.

- [0007] Y. Li, et al., A7.663-TOPS 8.2-W Energy-efficient FPGA Accelerator for Binary Convolutional Neural Networks, Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Pages 290-291, Feb. 2017에 기재된 수법에서는, 서로 다른 종류의 연산기를 사용하여 레이어마다 비트 폭이 다른 CNN을 처리한다. 입력 레이어의 특징 데이터의 비트 폭이 8비트이고, 중간 레이어의 특징 데이터의 비트 폭이 2비트인 경우, 8비트의 데이터 전용의 콘볼루션 연산부와 2비트의 데이터 전용의 콘볼루션 연산부가 필요하다.
- [0008] 8비트의 데이터의 레이어와 2비트의 데이터의 레이어를 파이프라인 방식으로 처리하여 데이터를 병렬로 처리할 수 있지만, 레이어마다 콘볼루션 처리의 계산량이 다른 경우에는, 하드웨어의 이용 효율이 저하한다. 또한, 2비트와 8비트 사이의 비트 폭(4비트 등)을 갖는 특징 데이터를 처리할 경우, 이와 같은 비트 폭에 대한 콘볼루션 연산부가 존재하지 않기 때문에, 8비트 데이터 전용의 콘볼루션 연산부를 이용해야 하므로, 효율이 저하한다.
- [0009] K. Lee, et al., A 502-GOPS and 0.984-mW Dual-Mode Intelligent ADAS SoC With Real-Time Semiglobal Matching and Intention Prediction for Smart Automotive Black Box System, IEEE Journal of Solid-State Circuits, Vol.52, No.1, Pages 139-150, Jan. 2017에 기재된 수법에서는, 복수 종류의 비트 폭을 갖는 특징 데이터를 처리가능한 SIMD(single instruction multiple data) 구성을 갖는 RNN(Reccurent Neural Network) 전용 하드웨어가 제안되어 있다. 같은 하드웨어를 이용하여 8비트, 16비트 및 32비트의 데이터를 처리 가능하지만, 병렬로 출력된 데이터의 총합을 계산할 때에, 데이터가 일단 메모리에 유지한 후 1회 SIMD의 코멘드를 실행할 필요가 있기 때문에, 처리 시간이 증가한다.

발명의 내용

해결하려는 과제

- [0010] 본 발명은, 복수의 계층의 네트워크 중에 복수 종류의 비트 폭을 갖는 데이터 세트들이 존재하여도 효율적인 처리를 실현하기 위한 기술을 제공한다.

과제의 해결 수단

- [0011] 본 발명의 제1 면에 따르면, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와, 상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC(multiply-accumulate) 연산을 실행하도록 구성된 복수의 MAC부와, 상기 복수의 MAC부에 의해 얻어진 복수의 MAC 연산 결과를 상기 시프트량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와, 상기 복수의 시프트 연산부에 의해 시프트된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부를 구비한 정보 처리장치가 제공된다.
- [0012] 본 발명의 제2 면에 따르면, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와, 상기 계층의 복수의 데이터를 상기 시프트량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와, 복수의 필터 계수와, 상기 복수의 시프트 연산부에 의해 시프트된 복수의 데이터에 대해, MAC 연산을 실행하도록 구성된 복수의 MAC부와, 상기 복수의 MAC부에 의해 계산된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부를 구비한 정보 처리장치가 제공된다.
- [0013] 본 발명의 제3 면에 따르면, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하는 단계와, 상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC 연산을 실행하는 단계와, 복수의 상기 MAC 연산의 결과를 상기 시프트량에 근거하여 시프트하는 단계와, 시프트된 상기 복수의 MAC 연산 결과의 총합을 계산하는 단계를 포함하는 정보 처리방법이 제공된다.
- [0014] 본 발명의 제4 면에 따르면, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하는 단계와, 상기 계층의 복수의 데이터를 상기 시프트량에 근거하여 시프트하는 단계와, 복수의 필터 계수와 시프트된 상기 복수의 데이터에 대해, MAC 연산을 실행하는 단계와, 계산된 복수의 MAC 연산 결과의 총합을 계산하는 단계를 포함하는 정보 처리방법이 제공된다.
- [0015] 본 발명의 제5 면에 따르면, 컴퓨터를, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트량을 설정하도록 구성된 제어부와, 상기 계층의 복수의 데이터와 복수의 필터 계수에 대해 MAC 연산을 실행하도록 구성된 복수의 MAC부와, 상기 복수의 MAC부에 의해 얻어진 복수의 MAC 연산 결과를 상기

시프트 량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와,

[0016] 상기 복수의 시프트 연산부에 의해 시프트된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부로서 기능시키는 컴퓨터 프로그램을 기억하는 비일시적인 컴퓨터 판독가능한 기억매체가 제공된다.

[0017] 본 발명의 제6 면에 따르면, 컴퓨터를, 복수의 계층을 갖는 네트워크의 각각의 계층에 대해, 데이터의 비트 폭에 근거하여 시프트 량을 설정하도록 구성된 제어부와, 상기 계층의 복수의 데이터를 상기 시프트 량에 근거하여 시프트하도록 구성된 복수의 시프트 연산부와, 복수의 필터 계수와, 상기 복수의 시프트 연산부에 의해 시프트된 복수의 데이터에 대해, MAC 연산을 실행하도록 구성된 복수의 MAC부와, 상기 복수의 MAC부에 의해 계산된 복수의 MAC 연산 결과의 총합을 계산하도록 구성된 가산부로서 기능시키는 컴퓨터 프로그램을 기억하는 비일시적인 컴퓨터 판독가능한 기억매체가 제공된다.

[0018] 본 발명의 또 다른 특징은 (첨부도면을 참조하는) 이하의 실시형태의 설명으로부터 명백해질 것이다.

도면의 간단한 설명

[0019] 도1은, 데이터 처리의 흐름도이다.

도2는, 처리 대상 네트워크의 구성예를 도시한 도면이다.

도3은, 정보 처리장치의 하드웨어 구성예를 나타낸 블록도이다.

도4는, 데이터 처리부(305)의 구성예를 나타낸 블록도이다.

도5는, 곱합연산부(402)와 시프트 연산부(403)의 구성예를 나타낸 블록도이다.

도6a 내지 도6c는, 데이터와 처리 시간의 관계를 나타낸 표이다.

도7은, 데이터 처리부(305)의 구성예를 나타낸 블록도이다.

도8은, 곱합연산부(702)와 시프트 연산부(701)의 구성예를 나타낸 블록도이다.

도9는, 데이터 처리의 흐름도이다.

도10a 내지 도10c는, 8비트 특징 데이터, 2비트 특징 데이터 및 4비트 특징 데이터의 예를 각각 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 이하, 첨부도면을 참조하여, 본 발명의 실시형태에 대해 설명한다. 이때, 이하 설명하는 실시형태는, 본 발명을 구체적으로 실시한 경우의 일례를 나타낸 것으로, 청구범위에 기재한 구성의 구체적인 실시형태의 1개이다.

[0021] [제1 실시형태]

[0022] 우선, 본 실시형태에 따른 정보 처리장치의 하드웨어 구성예에 대해, 도3의 블록도를 참조하여 설명한다. 정보 처리장치로서, PC(퍼스널컴퓨터), 태블릿형 단말장치, 스마트 폰 등의 컴퓨터 장치를 적용할 수 있다. 또한, 이 정보 처리장치는, 이들의 기기 등에 임베디드되는 임베디드 기기이어도 된다.

[0023] 입력부(301)는, 키보드, 마우스, 터치패널 등의 유저 인터페이스에 의해 구성되어 있고, 유저가 조작할 때 각종의 지시를 CPU(306)에 입력할 수 있다.

[0024] 데이터 보존부(302)는, 하드디스크 드라이브 장치 등의 대용량 정보 기억장치다. 데이터 보존부(302)에는, OS(오퍼레이팅 시스템), CPU(306)이 실행하는 각종의 컴퓨터 프로그램, CPU(306)이 각종의 처리를 실행할 때에 사용하는 데이터 등, 정보 처리장치에서 사용될 각종의 정보가 보존되어 있다. 데이터 보존부(302)에 보존되어 있는 데이터에는, 화상처리부(309)에 의한 처리 대상이 되는 화상이 포함되어 있다. 이때, 이하에서 "기지의 정보"로서 설명하는 정보도, 이 데이터 보존부(302)에 보존되어 있다. 데이터 보존부(302)에 보존되어 있는 컴퓨터 프로그램과 데이터는, CPU(306), 데이터 처리부(305) 및 화상처리부(309)에 의해 RAM(308) 등에 로드되고, CPU(306), 데이터 처리부(305) 및 화상처리부(309)의 처리 대상이 된다.

[0025] 이때, 데이터 보존부(302)는, 기억매체(예를 들어, 플렉시블 디스크, CD-ROM, CD-R, DVD, 메모리카드, CF 카드, 스마트 미디어, SD 카드, 메모리 스틱, xD 픽처 카드, USB 메모리 등)이어도 된다. 이 경우, 정보 처리장치는, 이러한 기억매체에 대한 정보의 판독 및 기록을 행하는 장치를 가질 필요가 있다.

- [0026] 통신부(303)는, 외부의 기기와 데이터 통신을 행하기 위한 통신 인터페이스로서 기능하는 것이다. 정보 처리장치에 있어서 처리의 실행에 필요한 정보를 이 통신부(303)가 외부에서 취득하도록 해도 된다. 통신부(303)는, 정보 처리장치가 행한 처리의 결과를 외부의 기기에 송신해도 된다.
- [0027] 표시부(304)는, 액정 화면이나 터치패널 화면 등에 의해 구성되어 있고, CPU(306), 데이터 처리부(305)나 화상 처리부(309)에 의해 얻어진 처리 결과를 표시하기 위해 화상, 문자 등을 표시할 수 있다. 이때, 표시부(304)는, 프로젝터 장치 등의 투영 장치이어도 된다. 입력부(301)와 표시부(304)를 일체화시켜서 터치스크린 장치 등의, 지시 입력의 접수 기능과 표시 기능을 갖는 장치를 형성해도 된다.
- [0028] 데이터 처리부(305)는, 화상처리부(309)에 의해 RAM(308)에 기록된 화상을 사용해서 도1의 흐름도에 따른 처리를 실행함으로써 CNN의 계산을 행하고, 얻어진 계산 결과를 데이터 보존부(302), RAM(308) 등에 출력한다. 이때, 데이터 처리부(305)의 처리 대상이 되는 화상은, 화상처리부(309)에 의해 RAM(308)에 기록된 화상에 한정되지 않고, 예를 들면, 다른 장치에 의해 입력된 화상이어도 된다. 데이터 처리부(305)에 대해서는 도4를 참조해서 후술한다.
- [0029] CPU(306)은, ROM(307)이나 RAM(308)에 격납되어 있는 컴퓨터 프로그램과 데이터를 사용해서 각종의 처리를 실행한다. 이에 따라, CPU(306)은 정보 처리장치 전체의 동작 제어를 행할 수 있다.
- [0030] ROM(307)에는, 정보 처리장치의 설정 데이터, 기동 프로그램 등의 고쳐쓰기가 불필요한 정보가 격납되어 있다. RAM(308)은, 데이터 보존부(302)와 ROM(307)으로부터 로드된 컴퓨터 프로그램과 데이터, 통신부(303)가 외부에서 수신한 정보 등을 격납하기 위한 에어리어를 가진다. RAM(308)은, CPU(306), 데이터 처리부(305) 및 화상처리부(309)가 각종의 처리를 실행할 때에 사용하는 워크 에어리어를 가진다. 이렇게 RAM(308)은 각종의 에어리어를 적절히 제공할 수 있다.
- [0031] 화상처리부(309)는, CPU(306)로부터의 지시하에서, 데이터 보존부(302)에 보존되어 있는 화상을 판독하고, 이 화상의 각 화소의 화소값의 레인지 조정을 행한 후, 이 화상을 RAM(308)에 기록한다.
- [0032] 상기한 입력부(301), 데이터 보존부(302), 통신부(303), 표시부(304), 데이터 처리부(305), CPU(306), ROM(307), RAM(308) 및 화상처리부(309)는 모두 버스(310)에 접속되어 있다.
- [0033] 이때, 정보 처리장치의 하드웨어 구성은 도3에 나타난 구성에 한정되지 않는다. 예를 들면, 도3의 구성을 복수의 장치로 실현시켜도 된다. 또한, 입력부(301), 데이터 보존부(302) 및 표시부(304) 등의 기기를 정보 처리장치가 항상 포함할 필요는 없으며, 이들 기기를 정보 처리장치와 통신로를 거쳐 접속하도록 설정해도 된다.
- [0034] 또한, RAM(308)에 격납하는 것으로서 설명하는 정보의 일부 혹은 전부를 데이터 보존부(302)에 격납해도 되고, 데이터 보존부(302)에 격납하는 것으로서 설명하는 정보의 일부 혹은 전부를 RAM(308)에 격납해도 된다. 또는, RAM(308)의 일부를 데이터 보존부(302)로서 사용하도록 설정하거나, 통신부(303)의 통신 상대의 기기의 기억장치를, 통신부(303)를 거쳐 데이터 보존부로서 이용하도록 가상적으로 구성해도 된다.
- [0035] 또한, 도3에서는 단지 1개의 CPU(306)가 도시되어 있지만, 정보 처리장치가 갖는 CPU(306)의 수는 1개에 한정되지 않고, 정보 처리장치에 복수의 CPU가 포함되어도 된다. 또한, 데이터 처리부(305)와 화상처리부(309)는 하드웨어로서 실장해도 되고, 컴퓨터 프로그램으로서 실장해도 된다. 후자의 경우, 이 컴퓨터 프로그램은 데이터 보존부(302)에 격납되고, CPU(306)이 대응하는 컴퓨터 프로그램을 실행함으로써 데이터 처리부(305)와 화상처리부(309)의 기능을 실현시키게 된다.
- [0036] 이때, CPU(306)은, 데이터 처리부(305)의 처리의 결과에 근거하여, 통신부(303)나 데이터 보존부(302)로부터 취득하는 동화상의 각 프레임에 대한 화상처리 및/또는 화상 인식을 행한다. CPU(306)에 의한 화상처리나 화상 인식의 결과는, RAM(308)이나 데이터 보존부(302)에 보존되거나, 또는 통신부(303)를 거쳐 외부의 기기에 대하여 출력된다. 또한, CPU(306)에 의한 화상처리나 화상 인식의 결과는, 표시부(304)에 화상이나 문자로서 표시되거나, 정보 처리장치가 음성 출력 기능을 가질 경우에는, 음성으로서 출력해도 된다.
- [0037] <처리 대상 네트워크>
- [0038] 본 실시예에서는 처리 대상 네트워크로서 CNN을 사용한다. 도2는 처리 대상 네트워크의 구성예를 나타낸 것이다. 도2의 처리 대상 네트워크의 상세에 대해서는 상기한 것과 같다. 이때, 처리 대상 네트워크의 각 MAC 연산의 계산량, 각 특징 화상의 사이즈, 특징 화상의 매수, 각 특징 화상의 비트 수 등의 정보는 데이터 보존부(302) 등에 보존되어 있다.

- [0039] 도2에 도시된 처리 대상 네트워크의 레이어의 수는 4(레이어 1 내지 4)이며, 각 레이어 중에 4개의 특징 화상이 존재한다. 상기한 바와 같이, 특징 화상 (i,j)은, 레이어 i의 j번째의 특징 화상을 나타낸다. 또한, 레이어에 따라, 레이어의 특징 화상의 비트 폭이 다르다. 레이어 1의 각 특징 화상의 비트 폭은 8비트이고, 레이어 2의 각 특징 화상의 비트 폭은 2비트이고, 레이어 3의 각 특징 화상의 비트 폭은 4비트이고, 레이어 4의 각 특징 화상의 비트 폭은 8비트다. 최초의 레이어(레이어 1)와 최후의 레이어(레이어 4)는 입출력 화상의 정보를 유지하기 때문에, 최초의 레이어와 최후의 레이어의 비트 폭이 중간 레이어들(레이어 2 및 레이어 3)의 비트 폭(2비트 및 4비트)보다 커지는 경향이 있다. 각 특징 화상은 복수의 화소(특징 데이터)로 구성되어 있다.
- [0040] 이하, 데이터 처리부(305)에 의해 실행되는 레이어 1 내지 4의 각각의 특징 화상의 계산(생성)에 대해 설명한다. 필터 계수와, 레이어 1의 8비트의 특징 화상 (1,1), (1, 2), (1, 3), (1, 4)을 사용해서 상기한 수학적 식 1에 따른 MAC 연산을 행한다. 그후, 이 MAC 연산의 결과로서, 레이어 2의 2비트의 특징 화상 (2, 1), (2, 2), (2, 3), (2, 4)을 생성한다.
- [0041] 다음에, 레이어 2의 2비트의 특징 화상 (2, 1), (2, 2), (2,3), (2, 4)과, 필터 계수를 사용해서 상기한 수학적 식 1에 따른 MAC 연산을 행한다. 그후, 이 MAC 연산의 결과로서, 레이어 3의 4비트의 특징 화상 (3,1), (3,2), (3,3), (3,4)을 생성한다.
- [0042] 그후, 레이어 3의 4비트의 특징 화상 (3, 1), (3, 2), (3,3), (3, 4)과, 필터 계수를 사용해서 상기한 수학적 식 1에 따른 MAC 연산을 행한다. 그후, 이 MAC 연산의 결과로써, 레이어 4의 8비트의 특징 화상 (4,1), (4,2), (4,3), (4,4)을 생성한다.
- [0043] <데이터 처리부(305)의 구성예>
- [0044] 데이터 처리부(305)의 구성예를 도4에 나타낸다. 데이터 메모리(401)는, 각 레이어의 각각의 특징 화상의 특징 데이터를 유지하고 있고, 계수 메모리(404)는 필터 계수를 유지하고 있다. 곱합연산부(402)는, 계수 메모리(404)에 유지되어 있는 필터 계수와, 데이터 메모리(401)에 유지되어 있는 특징 데이터를 사용하여 MAC 연산을 행함으로써, 각각의 MAC 연산 결과를 계산한다. 시프트 연산부(403)는, 곱합연산부(402)에 의해 얻어진 각각의 MAC 연산 결과를 시프트하고, 가산부(406)는, 시프트된 복수의 MAC 연산 결과를 가산해서 "시프트된 MAC 연산 결과의 총합"을 구한다. 처리부(407)는, 가산부(406)가 구한 "시프트된 MAC 연산 결과의 총합"에 근거하여, 활성화/풀링의 처리 결과를 계산하고, 이 계산한 활성화/풀링의 처리 결과를 데이터 메모리(401)에 격납한다. 제어부(405)는, 데이터 처리부(305) 전체의 동작 제어를 행한다.
- [0045] 데이터 처리부(305)에 의한 데이터 처리에 대해 도1의 흐름도에 따라서 설명한다. 스텝 S101에서는, 제어부(405)는, 복수개의 입력 특징 화상의 특징 데이터와 필터 계수를 RAM(308)로부터 판독하고, 이 판독한 특징 데이터를 데이터 메모리(401)에 격납하고, 이 판독한 필터 계수를 계수 메모리(404)에 격납한다.
- [0046] 스텝 S102에서는, 제어부(405)는, 레이어의 루프를 개시하고, 미처리의 레이어 중 1개를 처리 대상 레이어로 설정한다. 이 예에서는 레이어 1 내지 4가 순차적으로 처리 대상으로 설정되기 때문에, 처음에는 레이어 1이 처리 대상 레이어가 된다.
- [0047] 스텝 S103에서는, 제어부(405)는, 시프트 연산부(403)의 시프트 량을 정의하는 시프트 파라미터를 레이어 정보에 따라 설정한다. 스텝 S104에서는, 제어부(405)는, 출력 특징 화상의 루프를 개시하고, 출력 특징 데이터를 순차적으로 계산한다. 스텝 S105에서는, 제어부(405)는, 가산부(406)에 보존되어 있는 MAC 연산 결과를 초기화해서 MAC 연산 결과를 제로로 설정한다. 가산부(406)는 총합 계산의 기능을 갖추고 있다.
- [0048] 스텝 S106에서는, 제어부(405)는, 입력 특징 화상의 루프를 개시하고, 입력 특징 데이터를 순차적으로 처리한다. 스텝 S107에서는, 제어부(405)의 제어하에서, 곱합연산부(402) 및 시프트 연산부(403)는 전술한 MAC 연산 및 시프트 연산을 각각 행한다. 스텝 S107의 처리의 상세(스텝 S115 내지 스텝 S117)에 대해서는 후술한다.
- [0049] 스텝 S108에서는, 가산부(406)는, 복수의 MAC 연산 결과를 가산하여 "시프트된 MAC 연산 결과의 총합"을 구한다. 스텝 S109에서는, 제어부(405)는, 입력 특징 화상의 루프의 종료 판정을 한다. 모든 입력 특징 화상의 처리가 완료한 경우에는 스텝 S110으로 처리를 진행한다. 그렇지 않은 경우에는, 스텝 S107로 처리가 되돌아가, 다음의 미처리의 입력 특징 화상의 처리를 개시한다. 스텝 S110에서는, 처리부(407)는, 가산부(406)가 구한 "시프트된 MAC 연산 결과의 총합"에 근거하여, 이하의 수학적식에 따라서 활성화 처리 결과를 계산한다.

수학식 2

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

[0050]

[0051]

이 경우, $f(\cdot)$ 는 활성화 함수이며, x 는 입력 데이터다. 이 예에서는 ReLU(Rectified Linear Unit)을 사용해서 활성화 함수를 실현했지만, 본 발명은 ReLU에 한정되지 않고, 활성화 함수가 다른 비선형의 함수 또는 양자화 함수에 의해 실현되어도 된다. 이때, 필요에 따라 활성화 처리 결과의 비트 폭이 조정된다.

[0052]

스텝 S111에서는, 처리부(407)는 레이어의 정보에 따라, 활성화 처리 결과에 근거하여 풀링 처리를 행함으로써, 활성화/풀링의 처리 결과를 계산한다. 스텝 S112에서는, 처리부(407)는, 스텝 S111에서 계산한 활성화/풀링의 처리 결과를, 다음 레이어의 특징 화상으로서 데이터 메모리(401)에 격납한다.

[0053]

스텝 S113에서는, 제어부(405)는, 출력 특징 화상의 루프의 종료 판정을 한다. 모든 출력 특징 화상의 처리가 완료한 경우에는 스텝 S114로 처리를 진행하고, 그렇지 않은 경우에는, 스텝 S105로 처리가 되돌아가, 미처리의 출력 특징 화상의 처리를 개시한다.

[0054]

스텝 S114에서는, 제어부(405)는 레이어의 루프의 종료 판정을 한다. 모든 레이어의 처리가 완료한 경우, 도1의 흐름도에 따른 처리를 종료한다. 그렇지 않은 경우에는, 스텝 S103으로 처리가 되돌아가, 미처리의 레이어의 처리를 개시한다.

[0055]

<MAC 연산 및 시프트 연산>

[0056]

스텝 S107의 MAC 연산 및 시프트 연산(스텝 S115 내지 스텝 S117)에 대해 설명한다. 스텝 S115에서는, 제어부(405)는, 데이터 메모리(401)로부터 특징 데이터를 판독해서 판독한 특징 데이터를 곱합연산부(402)에 전송하는 것 이외에, 계수 메모리(404)로부터 필터 계수를 판독해서 판독한 필터 계수를 곱합연산부(402)에 전송한다. 필터 계수의 개수와 전송 횟수는 특징 데이터의 비트 폭에 따라 다르다.

[0057]

스텝 S116에서는, 곱합연산부(402)는, 특징 데이터와 필터 계수에 근거하여 MAC 연산 결과를 계산한다. 스텝 S117에서는, 시프트 연산부(403)는, 스텝 S116에서 얻어진 MAC 연산 결과를, 스텝 S103에서 설정된 시프트 파라미터가 표시하는 시프트 량에 근거하여 시프트한다.

[0058]

<비트 폭이 다른 경우의 상세 설명>

[0059]

본 실시형태에서는 비트 폭이 다른 데이터를 처리할 수 있다. 도 10a에는, 8비트의 특징 데이터를 처리할 경우에 있어서의 곱합연산부(402), 시프트 연산부(403) 및 가산부(406)의 동작을 나타내고 있다. 도 10b에는, 2비트의 특징 데이터를 처리할 경우에 있어서의 곱합연산부(402), 시프트 연산부(403) 및 가산부(406)의 동작을 나타내고 있다.

[0060]

특징 데이터가 8비트인 경우, 도 10a에 나타내는 것과 같이, 곱합연산부(402)는 8비트의 특징 데이터 1001(값: 234)을 2비트의 데이터(2비트 데이터)의 세트로 분할한다. 곱합연산부(402)는, 이 분할로부터 얻어진 4개의 세트의 2비트 데이터(값: 2, 2, 2, 3)와 공통의 필터 계수를 사용해서 4개의 MAC 연산 결과를 계산하고, 시프트 연산부(403)는, 4개의 시프트 파라미터에 근거하여 4개의 MAC 연산 결과를 시프트한다. 그후, 가산부(406)는, 시프트된 4개의 MAC 연산 결과를 가산해서 1개의 특징 데이터의 세트(8비트의 입력 특징 데이터의 MAC 연산 결과)를 계산한다. 이렇게, 데이터 처리부(305)는 1개 세트의 8비트의 입력 특징 데이터를 처리할 수 있다.

[0061]

특징 데이터가 2비트인 경우, 도 10b에 나타내는 것과 같이, 곱합연산부(402)는 4개의 세트의 2비트 데이터 1002(값: 2, 2, 2, 3)와 4개의 필터계수를 사용하여, 4개의 MAC 연산결과를 계산한다. 시프트 연산부(403)는, 1개의 시프트 파라미터에 근거하여 4개의 MAC 연산 결과를 시프트한다. 시프트 파라미터가 제로이기 때문에, MAC 연산 결과의 상태는 시프트 연산 전후에 동일하다. 그후, 가산부(406)는, MAC 연산 결과를 가산해서 1개의 세트의 특징 데이터(4개의 2비트의 입력 특징 데이터 세트들의 MAC 연산 결과의 총합)를 계산한다. 이렇게, 데이터 처리부(305)는 4개의 세트의 2비트의 입력 특징 데이터를 병렬로 처리할 수 있다.

[0062]

입력 특징 화상의 매수를 M 으로 하고, 필터 사이즈를 1×1 로 한다. 필터 사이즈가 1화소이며, 변수 x 및 y 의 값

은 상수이기 때문에, $O_{i,j}(n)$ 은 $I_{i,j}(n)$ 을 사용해서 계산한다. MAC 연산의 계산(수학식 1)이 이하와 같이 간략화 된다.

수학식 3

$$O(n) = \sum_{m=1}^M (I(m) \times C(m, n))$$

[0063]

[0064] 필터 사이즈가 1×1 보다 큰 경우, 곱합연산부(402)은 필터 계수와 입력 특징 데이터의 각 컨볼루션 연산 결과를 계산하지만, 필터 사이즈가 1×1 인 경우, 곱합연산부(402)는 $I(m)$ 와 $C(m, n)$ 의 곱을 계산한다.

[0065] 비트 폭이 α 비트인 특징 처리 데이터와 비트 폭이 β 비트인 특징 처리 데이터의, 2종류의 특징 처리 대상 데이터가 존재하는 것으로 가정한다. 도 4에 나타난 곱합연산부(402)는 MAC 연산 결과를 계산하는 P개의 α 비트의 데이터의 MAC 연산 유닛을 포함하고, 시프트 연산부(403)는 시프트 연산 결과를 계산하는 P개의 α 비트의 데이터의 시프트 연산 유닛을 포함한다. α , β , P는 이하의 조건을 만족하고 있다.

수학식 4

$$\beta = \alpha \times P$$

[0066]

[0067] 입력 특징 데이터 $I'_{(\beta)}$ 의 비트 폭이 β 비트인 경우, 가산기(406)의 출력은 수학식 6, 수학식 7, 수학식 8의 전제조건에 근거하여 이하의 수학식 5로 표시된다. n매체의 출력 화상의 MAC 연산 결과 $O(n)$ 는 다음과 같이 주어진다.

수학식 5

$$O(n) = \sum_{m=1}^M \sum_{p=1}^P [(I_{(\alpha),p}(m) \times C_p(m, n)) \times 2^{S(p)}]$$

[0068]

[0069] 이때, $I_{(\alpha),p}(m)$ 는 α 비트의 데이터의 MAC 연산 유닛의 입력 데이터이고, $C_p(m, n)$ 은 필터계수이며, $S(p)$ 는 시프트 파라미터이다. 변수 m 은, α 비트의 입력 특징 화상 그룹(1 그룹=P매의 화상)의 번호(곱합연산부(402)의 처리번호)이며, 변수 p 는 MAC 연산 유닛의 번호 및 시프트 연산 유닛의 번호이며, 변수 n 은 출력 특징 화상의 번호이다. 시프트 연산은 2의 거듭제곱에 의한 처리로 표현된다.

[0070] 필터계수 $C_p(m, n)$ 은 수학식 6에 나타난 것과 같이, m매체의 β 비트의 특징 화상에 대응하고 있는 필터 계수 $C'(m, n)$ 이다. α 비트의 입력 특징 화상 그룹에 대해 공통의 필터 계수가 사용되기 때문에, 변수 p 를 생략할 수 있다. P개의 MAC 연산 유닛에 병렬로 공급하는 필터 계수의 개수는 1개이며, 전송 횟수는 1회이다.

수학식 6

$$C_p(m, n) = \hat{C}(m, n)$$

[0071]

[0072] 이 경우, 입력 데이터 $I'_{(\beta)}$ 를 P개의 세트의 α 비트의 데이터 $I_{(\alpha),p}(m)$ 로 분할한다. 시프트 파라미터 $S(p)$ 의 값은 다음과 같이 MAC 연산 유닛의 번호 p와 분할된 데이터의 비트 폭 α 에 근거하여 계산한다.

수학식 7

$$S(p) = \alpha \times (p - 1)$$

[0073]

[0074] β 비트의 입력 특징 데이터 $I'_{(\beta)}$ 는 다음과 같이 분할된 P개의 세트의 α 비트의 데이터 $I_{(\alpha),p}(m)$ 로 표현된다.

수학식 8

$$\sum_{p=1}^P I_{(\alpha),p}(m) \times 2^{S(p)} = \hat{I}_{(\beta)}(m)$$

[0075]

[0076] 이 경우, 수학식 6, 수학식 7 및 수학식 8을 수학식 5에 대입하면, 출력 데이터 $O(n)$ 의 수학식은 이하와 같이 된다.

수학식 9

$$O(n) = \sum_{m=1}^M \sum_{p=1}^P [(I_{(\alpha),p}(m) \times C_p(m, n)) \times 2^{\alpha \times (p-1)}] = \sum_{m=1}^M (\hat{I}_{(\beta)}(m) \times \hat{C}(m, n))$$

[0077]

[0078] 한편, 입력 특징 데이터 $I'_{(\alpha)}$ 의 비트 폭이 α 비트인 경우, 가산부(406)의 출력은 수학식 11, 수학식 12 및 수학식 13의 전제조건에 근거하여 이하의 수학식 10으로 표시된다. n매체의 출력 화상의 MAC 연산 결과 $O(n)$ 은 다음과 같이 주어진다.

수학식 10

$$O(n) = \sum_{m=1}^{\frac{M}{P}} \sum_{p=1}^P [(I_{(\alpha),p}(m) \times C_p(m, n)) \times 2^{S(p)}]$$

[0079]

[0080] 이때, $I_{(\alpha),p}(m)$ 은 α 비트의 데이터의 MAC 연산 유닛의 입력 데이터이고, $C_p(m, n)$ 은 필터 계수이고, $S(p)$ 는 시

프트 파라미터이다. 변수 m 은 α 비트의 입력 특징 화상 그룹(1 그룹= P 매의 화상)의 번호(곱합연산부(402)의 처리번호)이고, 변수 p 는 MAC 연산 유닛의 번호 및 시프트 연산 유닛의 번호이고, 변수 n 은 출력 특징 화상의 번호이다. 시프트 연산은 2의 거듭제곱에 의한 처리로 표현된다.

[0081] 필터계수 $C_p(m, n)$ 은 제 $\{(m-1) \times P + p\}$ 매째의 α 비트의 특징 화상에 대응하고 있는 필터 계수 $C'((m-1) \times P + p, n)$ 이다. MAC 연산 유닛의 번호 p 에 따라 필터 계수가 다르기 때문에, P 개의 MAC 연산 유닛에 병렬로 공급하는 필터 계수의 개수는 P 개이며, 전송 횟수는 P 회이다.

수학식 11

$$C_p(m, n) = \hat{C}((m-1) \times P + p, n)$$

[0082]

[0083] 입력 특징 데이터 $I'_{(\alpha)}$ 는 α 비트의 데이터의 MAC 연산 유닛의 입력 데이터 $I_{(\alpha), p}(m)$ 가 되고, 시프트 파라미터 $S(p)$ 의 값은 다음과 같이 항상 0이다.

수학식 12

$$S(p) = 0$$

[0084]

[0085] P 개의 세트의 α 비트의 입력 특징 데이터 $I'_{(\alpha)}$ 를 그대로 MAC 연산 유닛에 입력하지만, P 개의 세트의 입력 데이터는 각각 다른 특징 화상의 특징 데이터이다. 특징 화상의 번호는, MAC 연산 유닛의 번호 p 와, 시프트 연산 유닛의 개수 P 와, 곱합연산부(402)의 처리 번호 m 으로 이하의 수학식 13에 나타낸 것과 같이 표현된다.

수학식 13

$$I_{(\alpha), p}(m) = \hat{I}_{(\alpha)}((m-1) \times P + p)$$

[0086]

[0087] 수학식 11, 수학식 12 및 수학식 13을 수학식 10에 대입하면, 출력 데이터 $O(n)$ 의 수학식은 다음과 같이 된다.

수학식 14

$$O(n) = \sum_{m=1}^{\frac{M}{P}} \sum_{p=1}^P [\hat{I}_{(\alpha), p}((m-1) \times P + p) \times \hat{C}((m-1) \times P + p, n)]$$

[0088]

[0089] 시프트 파라미터 $S(p)$ 의 값과 필터 계수의 개수를 변경함으로써, 비트 폭이 α 비트인 특징 데이터 $I'_{(\alpha)}$ 와 비트 폭이 β 비트인 특징 데이터 $I'_{(\beta)}$ 를, 동일한 연산기(곱합연산부(402), 시프트 연산부(403) 및 가산부(406))를 사용하여 처리할 수 있다.

- [0090] <비트 폭이 다른 경우의 처리 예>
- [0091] 도5와 도10a 및 10b는, $P=4$, $\beta=8$, $\alpha=2$ 인 경우의 구성예이다. 곱합연산부(402)에 대한 입력 데이터의 비트 폭이 2비트이고, 시프트 연산부(403)에 대한 입력 데이터의 비트 폭이 6비트이고, 가산부(406)에 대한 입력 데이터의 비트 폭이 12비트다.
- [0092] 도6a 내지 도6c는 도 5에 도시된 하드웨어 구성을 사용해서 도2에 도시된 처리 대상 네트워크를 처리한 경우의 처리 시간의 예를 나타낸 것이다. 도6a와 도10a에 레이어 1(8비트의 데이터, 입력 특징 화상 매수 $M=4$)을 처리한 경우의 예를 나타낸다. 특징 화상 (1,1)의 특징 데이터 $I'_{(8)}(1)$ 는 8비트이고, 수학적 8에 근거하여 특징 데이터를 4개로 분할하여 얻어진 4개의 세트의 데이터 $I_{(2),1}(1)$ 내지 $I_{(2),4}(1)$ 를 곱합연산부(402)에 입력한다. 입력 특징 데이터 세트, 시프트 파라미터와, 필터 계수 $C(m,n)$ 를 사용하여 시프트 연산 결과를 계산하고, 계산된 시프트 연산 결과를 가산기(406)에 입력하여, 얻어진 결과에 초기값 제로가 가산된다. 계산 결과는 시프트 연산 결과로 설정되고, 가산부(406)에 의해 유지된다. 이 처리의 소요 시간은 1ms이다.
- [0093] 특징 화상 (1,2)의 특징 데이터 $I'_{(8)}(2)$ 는 8비트이며, 수학적 8에 근거하여 특징 데이터를 4개로 분할하여 얻어진 4개의 세트의 데이터 $I_{(2),1}(2)$ 내지 $I_{(2),4}(2)$ 를, 곱합연산부(402)에 입력한다. 입력 특징 데이터 세트, 시프트 파라미터와 필터 계수 $C(m,n)$ 를 사용해서 시프트 연산 결과를 계산하고, 계산된 시프트 연산 결과를 가산부(406)에 입력하여, 이전의 결과에 가산한다. 이 처리의 소요 시간은 1ms이다.
- [0094] 특징 화상 (1,2)과 유사한 방법으로, 특징 화상 (1,3)과 (1,4)를 순차적으로 처리하여, 시프트 연산 결과를 누적하고, 가산 결과를 계산한다. 이 처리의 소요 시간은 2ms이다. 최후에, 처리부(407)를 통해, 특징 화상 (2,1)의 특징 데이터를 출력한다. 4매의 특징 화상의 처리 시간은 4ms이다.
- [0095] 도6b와 도10b에 레이어 2(2비트의 데이터, 입력 특징 화상 매수 $M=4$)를 처리했을 경우의 예를 나타낸다. 특징 화상 (2,1) 내지 (2,4)의 각각의 특징 데이터의 세트 $I'_{(2)}(1)$ 내지 $I'_{(2)}(4)$ 는 2비트이며, 수학적 13에 근거하여 4개의 세트의 데이터 $I_{(2),1}(1)$ 내지 $I_{(2),4}(1)$ 를 곱합연산부(402)에 병렬로 입력한다. 입력 특징 데이터, 시프트 파라미터와 필터 계수 $C_p(m,n)$ 를 사용해서 시프트 연산 결과를 계산하여, 얻어진 결과를 가산부(406)에 입력하고, 초기값 제로와 가산하여, 계산 결과는 시프트 연산 결과가 된다. 최후에, 처리부(407)를 통해, 특징 화상 (3,1)의 특징 데이터를 출력한다. 4매의 특징 화상의 처리 시간은 1ms이다.
- [0096] 도6a 및 도6b와 도10a 및 도10b에 나타낸 것과 같이, 입력 특징 데이터가 8비트인 경우, 출력 데이터 당의 처리 시간은 4ms이며, 입력 특징 데이터가 2비트인 경우, 출력 데이터 당의 처리 시간은 1ms이다. 공통의 데이터 처리부(305)에 의해 비트 폭이 다른 데이터를 효율적으로 처리하는 것이 가능하다.
- [0097] [제2 실시형태]
- [0098] 이하에서는, 제1 실시형태와의 차분에 대해 설명한다. 이하에서 특별히 언급하지 않은 것은 제1 실시형태와 유사하다.
- [0099] <시프트 연산과 MAC 연산의 순서>
- [0100] 제1 실시형태에서는, MAC 연산 후에 시프트 연산을 행하는 예에 대해 설명하였다. 그러나, MAC 연산과 시프트 연산의 순서를 교체해도 동일한 처리 결과가 얻어질 수 있다. MAC 연산과 시프트 연산의 순서가 교체된 경우, 도1의 흐름도의 일부가 변경된다. 스텝 S107은 도9의 스텝 S901 내지 S903으로 변경된다.
- [0101] 도 7은 본 실시형태에 따른 데이터 처리부(305)의 구성예를 나타낸 것이다. 시프트 연산부(701)는, 데이터 메모리(401)에 격납되어 있는 특징 데이터를 시프트 파라미터에 근거하여 시프트하고, 곱합연산부(702)는, 시프트된 특징 데이터와 필터 계수에 근거하여 MAC 연산 결과를 계산한다.
- [0102] <시프트 연산 및 MAC 연산>
- [0103] 스텝 S107에서 행해지는 MAC 연산 및 시프트 연산(스텝 S901 내지 S903)에 대해 설명한다. 스텝 S901에서는, 제어부(704)는, 데이터 메모리(401)로부터 특징 데이터를 판독하고, 계수 메모리(703)로부터 필터 계수를 판독한다. 스텝 S902에서는, 시프트 연산부(701)는, 스텝 S103에서 설정된 시프트 파라미터에 근거하여, 특징 데이터를 시프트한다. 스텝 S903에서는, 곱합연산부(702)는, 시프트된 특징 데이터와 필터 계수에 근거하여 MAC 연산 결과를 계산한다.

[0104] <비트 폭이 다른 경우의 상세 설명>

[0105] 본 실시형태에서는, 시프트 연산부(701)는 시프트 연산 결과를 계산하는 P개의 α 비트의 데이터의 시프트 연산 유닛을 포함하고, 곱합연산부(702)는 MAC 연산 결과를 계산하는 P개의 α 비트의 데이터의 MAC 연산 유닛을 포함한다. 곱합연산부(702)의 출력은 이하의 수학적 식 15로 표시되고, 수학적 식 5에 나타난 시프트 연산부(403)의 출력과 등가이다.

수학적 식 15

$$O(n) = \sum_{m=1}^M \sum_{p=1}^P [(I_{(\alpha),p}(m) \times 2^{S(p)}) \times C_p(m, n)]$$

[0106]

[0107] 도8에, $P=4$, $\beta=8$ 및 $\alpha=2$ 인 경우의 예를 나타낸다. 시프트 연산부(701)에 대한 입력 데이터의 비트 폭이 2비트이고, 곱합연산부(702)에 대한 입력 데이터의 비트 폭이 8비트이며, 가산부(406)에 대한 입력 데이터의 비트 폭이 12비트이다. 시프트 연산부(701)의 회로 규모와 곱합연산부(702)의 회로 규모는 비트 폭의 차이에 의해 다르기 때문에, 시프트 연산부(701)(시프트 연산부(403))와 곱합연산부(702)(곱합연산부(402))의 순서를 교체함으로써, 전체적인 회로 규모를 삭감하는 것이 가능하다.

[0108] [제3 실시형태]

[0109] 제1 및 제2 실시형태에서는, 입력 특징 데이터의 비트 폭이 α 비트(각각의 MAC 연산 유닛의 비트 폭) 및 β 비트(각각의 MAC 연산 유닛의 비트 폭과 MAC 연산 유닛 수의 곱)인 예에 대해 설명하였다. 그러나, 본 발명은 이것들에 한정되지 않으며, α 및 β 이외의 비트 폭을 사용해도 된다.

[0110] <입력 특징 데이터의 비트 폭이 γ 비트인 경우>

[0111] 본 실시형태에서는, 비트 폭이 γ 비트인 입력 특징 데이터를 처리하는 것이 가능하다. 도10c에 특징 데이터가 4비트인 예를 나타낸다. 특징 데이터가 4비트인 경우, 도10c에 나타난 것과 같이, 곱합연산부(402)는 2개의 세트의 4비트의 특징 데이터 1003(값: 10, 14) 각각을 2비트씩 분할한다. 곱합연산부(402)는, 이 분할에 의해 얻어진 4개의 세트의 2비트의 데이터(값: 2, 2, 2, 3)와 2개의 필터 계수를 사용하여, 4개의 MAC 연산 결과를 계산한다. 시프트 연산부(403)는, 2개의 시프트 파라미터에 근거하여, 4개의 MAC 연산 결과를 시프트한다. 가산부(406)는, 시프트된 4개의 MAC 연산 결과를 가산하여 1개의 세트의 특징 데이터(2개의 세트의 4비트의 입력 특징 데이터의 MAC 연산 결과의 총합)를 계산한다. 이와 같이, 데이터 처리부(305)는, 2개의 세트의 4비트의 입력 특징 데이터를 병렬로 처리할 수 있다. γ 는 입력 특징 데이터의 비트 폭이고, γ 의 값은 β 의 값과 다르다. α , β 및 P의 정의는 1 실시형태와 같으며, γ , α 및 P'은 이하의 조건을 만족하고 있다.

수학적 식 16

$$\gamma = \alpha \times P'$$

[0112]

[0113] 이때, γ 는 β 보다 작고, P는 P'의 배수다. 입력 특징 데이터 $I'_{(\gamma)}$ 의 비트 폭이 γ 비트인 경우, 가산부(406)의 출력 데이터 $O(n)$ 는 수학적 식 18, 수학적 식 19 및 수학적 식 20의 전체조건에 근거한 이하의 수학적 식 17로 표시된다. n번째의 출력 특징 화상의 MAC 연산 결과 $O(n)$ 는 다음과 같이 주어진다.

수학식 17

$$O(n) = \sum_{m=1}^{\frac{MP}{P}} \sum_{q=1}^{\frac{P}{P}} \sum_{p=1}^P [(I_{(\alpha),p,q}(m) \times C_{p,q}(m,n)) \times 2^{S(p,q)}]$$

[0114]

[0115] 이때, $I_{(\alpha),p}(m)$ 는 α 비트의 데이터의 MAC 연산 유닛의 입력 데이터이고, $C_p(m,n)$ 은 필터 계수이고, $S(p)$ 는 시프트 파라미터이다. 변수 m 은 α 비트의 입력 특징 화상 그룹(1 그룹= P 개의 화상)의 번호(곱합연산부(402)의 처리번호)이다. MAC 연산 유닛은 P/P' 세트로 분할되고, 시프트 연산 유닛은 P/P' 세트로 분할되고, 변수 q 는 MAC 연산 유닛의 세트의 번호이다. 변수 p 는 세트 내의 MAC 연산 유닛 번호 및 시프트 연산 유닛의 번호이고, 변수 n 은 출력 특징 화상의 번호이다. 시프트 연산은 2의 거듭제곱에 의한 처리로 표현된다.

[0116] 필터 계수 $C_{p,q}(m,n)$ 은 제 $\{(m-1) \times P/P' + q\}$ 번째의 γ 비트의 특징 화상에 대응하고 있는 필터 계수 $C'((m-1) \times P/P' + q, n)$ 이다. 필터 계수는 MAC 연산 유닛의 세트 번호 q 에 근거하여 계산한다. 필터 계수의 일부가 공통되기 때문에, P 개의 MAC 연산 유닛에 병렬로 공급하는 필터 계수의 개수는 P/P' 개이며, 전송 횟수는 P/P' 회이다.

수학식 18

$$C_{p,q}(m,n) = C'((m-1) \times \frac{P}{P'} + q, n)$$

[0117]

[0118] 이 경우, 입력 특징 데이터 $I'_{(\gamma)}$ 를 P' 개 세트의 α 비트의 데이터 $I_{(\alpha),p}(m)$ 로 분할한다. 시프트 파라미터 $S(\cdot)$ 는 MAC 연산 유닛의 비트 폭 α 와 MAC 연산 유닛의 번호 p 에 근거하여 계산한다.

수학식 19

$$S(p,q) = \alpha \times (p-1)$$

[0119]

[0120] γ 비트의 입력 특징 데이터 $I'_{(\gamma)}$ 는 분할된 P' 개의 세트의 α 비트의 데이터 $I_{(\alpha),p,q}(m)$ 으로 표현된다.

수학식 20

$$\sum_{p=1}^P (I_{(\alpha),p,q}(m) \times 2^{S(p,q)}) = I_{(\gamma)}((m-1) \times \frac{P}{P'} + q)$$

[0121]

[0122] 수학식 18, 수학식 19 및 수학식 20을 수학식 17에 대입하면, 다음과 같은 출력 데이터 $O(n)$ 의 수학식이 얻어진

다.

수학식 21

$$O(n) = \sum_{m=1}^{\frac{M\beta}{P}} \sum_{q=1}^{\frac{P}{\beta}} \sum_{p=1}^{\beta} [(I_{(\alpha),p,q}(m) \times C_{p,q}(m,n)) \times 2^{\alpha \times (p-1)}]$$

$$= \sum_{m=1}^{\frac{M\beta}{P}} \sum_{q=1}^{\frac{P}{\beta}} [I_{(\gamma)}((m-1) \times \frac{P}{\beta} + q) \times C((m-1) \times \frac{P}{\beta} + q, n)]$$

[0123]

[0124] 시프트 파라미터 $S(p, q)$ 의 값과 필터 계수의 개수를 설정함으로써, 비트 폭이 γ 비트인 특징 데이터 $I'_{(\gamma)}$ 를, 제1 실시형태와 동일한 연산기(곱합연산부(402), 시프트 연산부(403) 및 가산부(406))를 사용하여 처리할 수 있다.

[0125] <비트 폭이 다른 경우의 처리 예>

[0126] 도 5와 도 10c에, $P=4$, $\beta=4$, $\alpha=2$ 인 경우의 구성예를 나타내고, 도 5의 하드웨어 구성을 사용하여 도 2의 처리 대상 네트워크를 처리한 경우의 처리 시간의 예를 도 6a 내지 도 6c에 나타낸다.

[0127] 도 6c와 도 10c에 $P'=2$ 및 $\gamma=4$ 이고, 레이어 3(4 비트의 데이터, 입력 특징 화상 매수 $M=4$)을 처리한 경우의 예를 나타낸다. 특징 화상 (3,1)과 (3,2)의 특징 데이터 $I'_{(4)}(1)$ 및 $I'_{(4)}(2)$ 각각은 4비트이고, 수학식 20에 근거하여 분할된 4개의 세트의 데이터 $I_{(2),1}(1)$ 내지 $I_{(2),4}(1)$ 를 곱합연산부(402)에 입력한다. 입력 특징 데이터, 시프트 파라미터와 필터 계수 $C(m,n)$ 를 사용하여, 시프트 연산 결과를 계산하고, 계산된 시프트 연산 결과를 가산부(406)에 입력하여, 계산된 결과에 초기값 제로가 가산된다. 계산 결과는 시프트 연산 결과가 되고, 가산부(406)에 의해 유지된다. 이 처리의 소요 시간은 1ms이다.

[0128] 특징 화상 (3',3)과 (3',4)의 특징 데이터 $I'_{(4)}(3)$ 및 $I'_{(4)}(4)$ 각각은 4비트이며, 수학식 20에 근거하여 분할된 4개의 세트의 데이터 $I_{(2),1}(2)$ 내지 $I_{(2),4}(2)$ 를, 곱합연산부(402)에 입력한다. 입력 특징 데이터, 시프트 파라미터와 필터 계수 $C(m,n)$ 를 사용하여, 시프트 연산 결과를 계산하고, 계산된 시프트 연산 결과를 가산부(406)에 입력하여, 이 결과가 이전의 결과에 가산된다. 이 처리의 소요 시간은 1ms이다. 최후에, 처리부(407)를 통해, 특징 화상 (4,1)의 특징 데이터를 출력한다. 4매의 특징 화상의 처리 시간은 2ms이다.

[0129] 이렇게, 본 실시형태는, 비트 폭이 α 비트(각각의 MAC 연산 유닛의 비트 폭) 또는 β 비트(각각의 MAC 연산 유닛의 비트 폭 α 와 MAC 연산 유닛의 수 P 의 곱)인 데이터 이외의 특징 데이터도 처리할 수 있기 때문에, 유연성이 높다고 하는 장점이 있다.

[0130] [제4 실시형태]

[0131] 제1 실시형태에서는, 처리부(407)에 의해 활성화 처리를 실행하는 예에 대해 설명했지만, 활성화 처리의 실행은 처리부(407)에 한정되지 않고, 다른 장치, 예를 들면, CPU(306)이 활성화 처리를 실행하도록 설정해도 된다. 이것은 다른 처리에 대해서도 마찬가지이며, 상기한 각 실시형태에서는, 상기한 다양한 처리의 주체의 일례를 나타낸 것에 지나지 않고, 상기한 각 실시형태에서 설명한 주체와는 다른 주체를 사용해도 된다.

[0132] 또한, 제1 실시형태에서는, 레이어 정보에 따라 활성화/풀링 처리를 실행하였다. 그러나, 경우에 따라서는, 활성화/풀링 처리를 생략해도 된다.

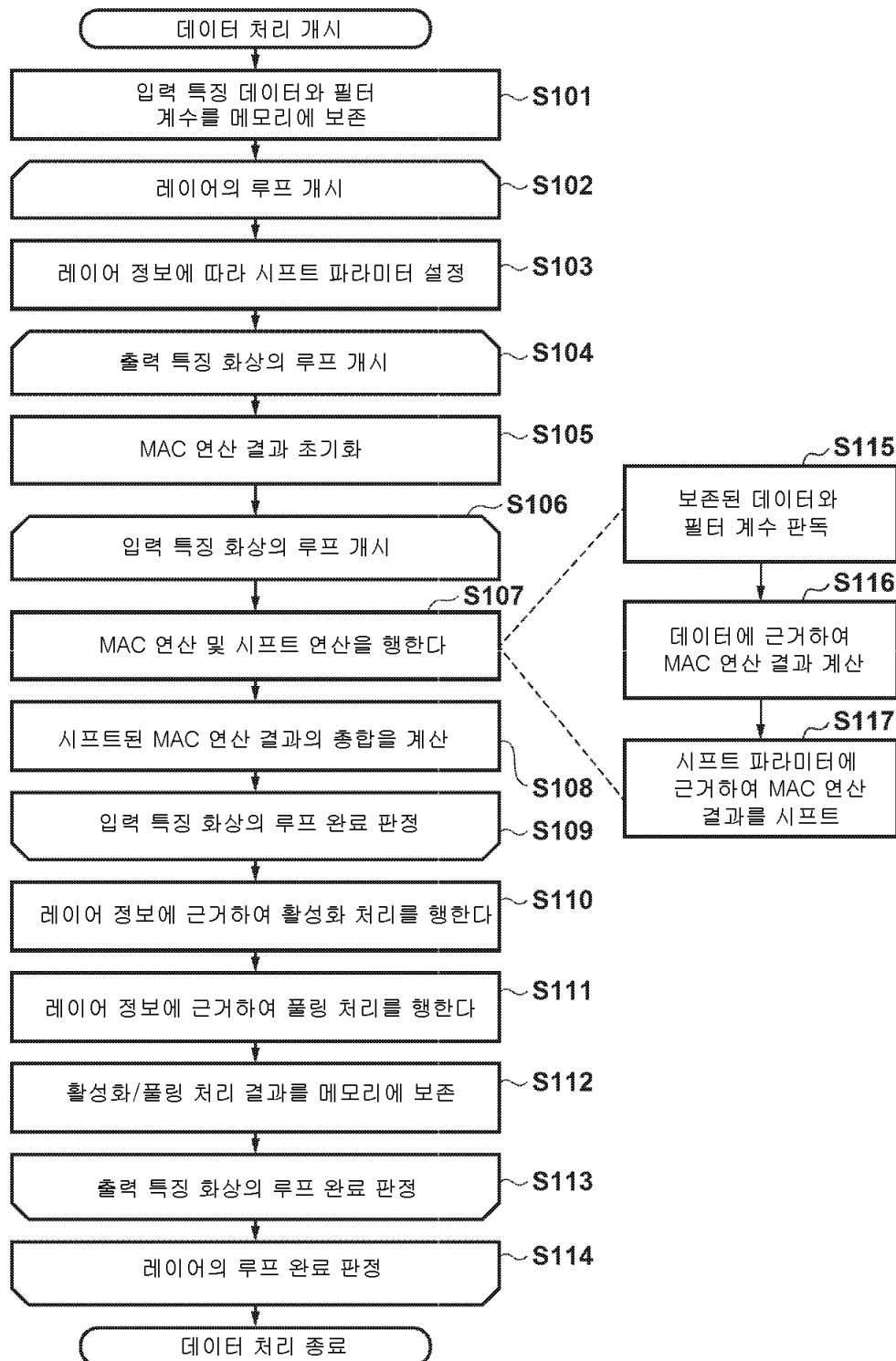
[0133] 또한, 제1 내지 3 실시형태에서는, 필터 사이즈(각 필터의 높이와 폭)가 1×1 인 케이스에 대해 설명했지만, 필터 사이즈는 1×1 에 한정되지 않고, 다른 사이즈이어도 된다. 상기한 각 실시형태의 설명에서 사용한 수치는 더욱 구체적인 설명을 행하기 위해서 사용된 예에 지나지 않고, 상기한 각 실시형태에서 설명한 수치에 수치를 한

정하려고 의도한 것이 아니다.

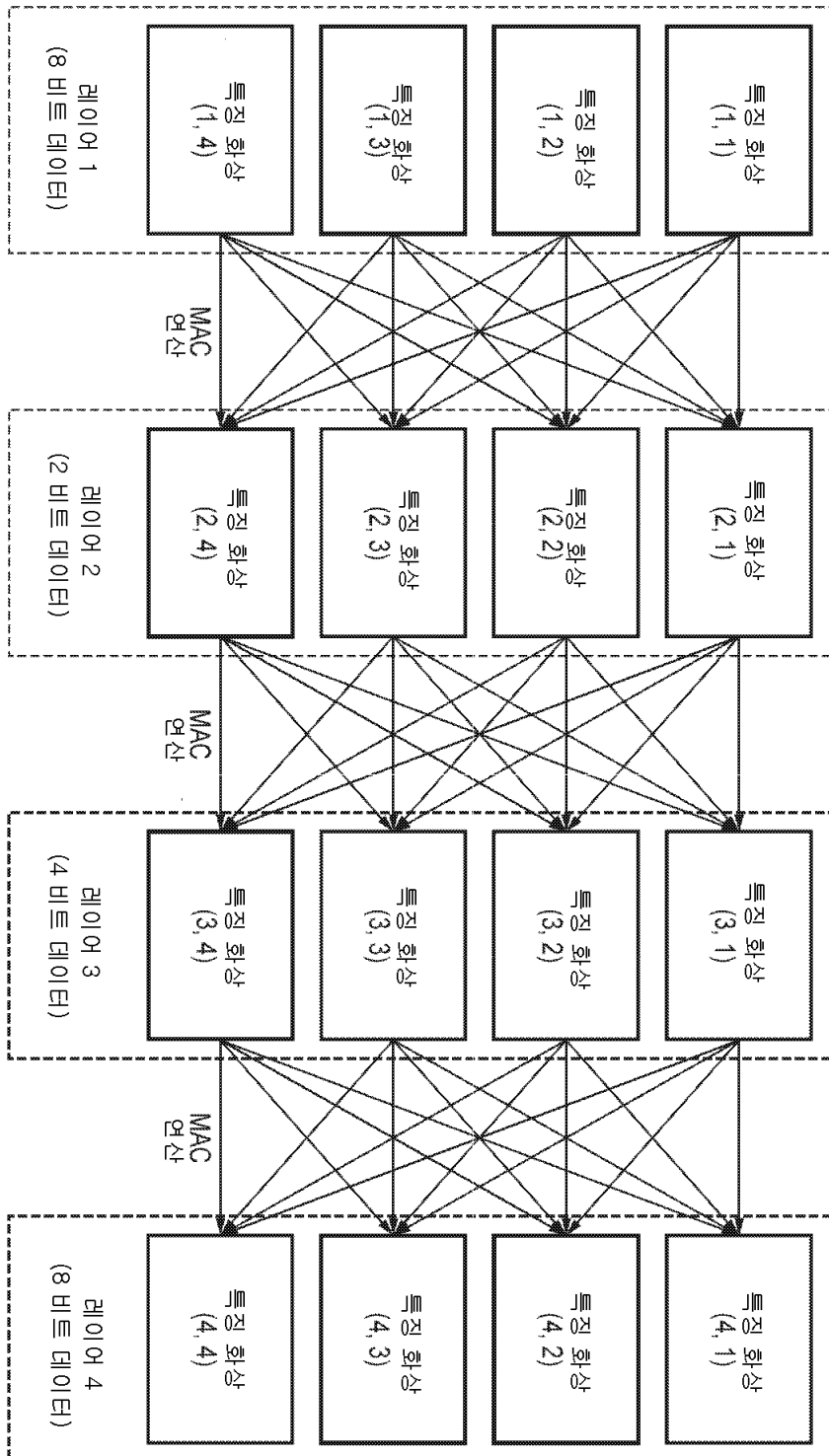
- [0134] 필터 사이즈가 작은 경우, 필터 계수를 유지하는 메모리(계수 메모리 404 또는 703)의 용량도 보다 작아질 수 있다는 장점이 있다. 필터의 폭과 필터 높이로서 설정되는 최소값은 1이다.
- [0135] 또한, 제1 내지 3 실시형태에서는, 입력 특징 화상의 매수를 M으로 설정하고, 출력 특징 화상의 매수를 N으로 설정하였다. 그러나, M, N에 적용가능한 수치는 특정한 수치에 한정되지 않는다. 이렇게, 상기한 다양한 종류의 변수에 적용가능한 수치는 특정한 수치에 한정되지 않는다.
- [0136] 또한, 제1 내지 3 실시형태에서는, 필터 계수를 계수 메모리 404 또는 703에 유지하고, 특징 데이터를 데이터 메모리(401)에 유지하였지만, 필터 계수와 특징 데이터를 유지하는 메모리는 특정한 메모리에 한정되지 않는다. 예를 들면, 필터 계수와 특징 데이터를 곱합연산부 402 또는 702에 포함된 메모리에 유지하거나, RAM(308)에 유지해도 된다.
- [0137] 또한, 각각의 필터 계수의 비트 폭은 특정한 비트 폭에 한정되지 않는다. 또한, 제1 내지 3 실시형태에서는, 처리 대상 네트워크로서 CNN을 사용했지만, 처리 대상 네트워크는 CNN에 한정되지 않고, RNN, MLP(다층 퍼셉트론) 등과 같은 다른 종류의 복수의 레이어가 계층적으로 접속되어 있는 네트워크이어도 된다.
- [0138] 기타 실시형태
- [0139] 본 발명의 실시형태는, 본 발명의 전술한 실시형태(들)의 1개 이상의 기능을 수행하기 위해 기억매체('비일시적인 컴퓨터 판독가능한 기억매체'로서 더 상세히 언급해도 된다)에 기록된 컴퓨터 실행가능한 명령(예를 들어, 1개 이상의 프로그램)을 판독하여 실행하거나 및/또는 전술한 실시예(들)의 1개 이상의 기능을 수행하는 1개 이상의 회로(예를 들어, 주문형 반도체 회로(ASIC)를 포함하는 시스템 또는 장치의 컴퓨터나, 예를 들면, 전술한 실시형태(들)의 1개 이상의 기능을 수행하기 위해 기억매체로부터 컴퓨터 실행가능한 명령을 판독하여 실행함으로써, 시스템 또는 장치의 컴퓨터에 의해 수행되는 방법에 의해 구현될 수도 있다. 컴퓨터는, 1개 이상의 중앙처리장치(CPU), 마이크로 처리장치(MPU) 또는 기타 회로를 구비하고, 별개의 컴퓨터들의 네트워크 또는 별개의 컴퓨터 프로세서들을 구비해도 된다. 컴퓨터 실행가능한 명령은, 예를 들어, 기억매체의 네트워크로부터 컴퓨터로 주어지기도 한다. 기록매체는, 예를 들면, 1개 이상의 하드디스크, 랜덤 액세스 메모리(RAM), 판독 전용 메모리(ROM), 분산 컴퓨팅 시스템의 스토리지, 광 디스크(컴팩트 디스크(CD), 디지털 다기능 디스크(DVD), 또는 블루레이 디스크(BD)TM 등), 플래시 메모리소자, 메모리 카드 등을 구비해도 된다.
- [0140] 본 발명은, 상기한 실시형태의 1개 이상의 기능을 실현하는 프로그램을, 네트워크 또는 기억매체를 개입하여 시스템 혹은 장치에 공급하고, 그 시스템 혹은 장치의 컴퓨터에 있어서 1개 이상의 프로세서가 프로그램을 읽어 실행하는 처리에서도 실행가능하다. 또한, 1개 이상의 기능을 실현하는 회로(예를 들어, ASIC)에 의해서도 실행가능하다.
- [0141] 예시적인 실시형태들을 참조하여 본 발명을 설명하였지만, 본 발명이 이러한 실시형태에 한정되지 않는다는 것은 자명하다. 이하의 청구범위의 보호범위는 가장 넓게 해석되어 모든 변형, 동등물 구조 및 기능을 포괄하여야 한다.

도면

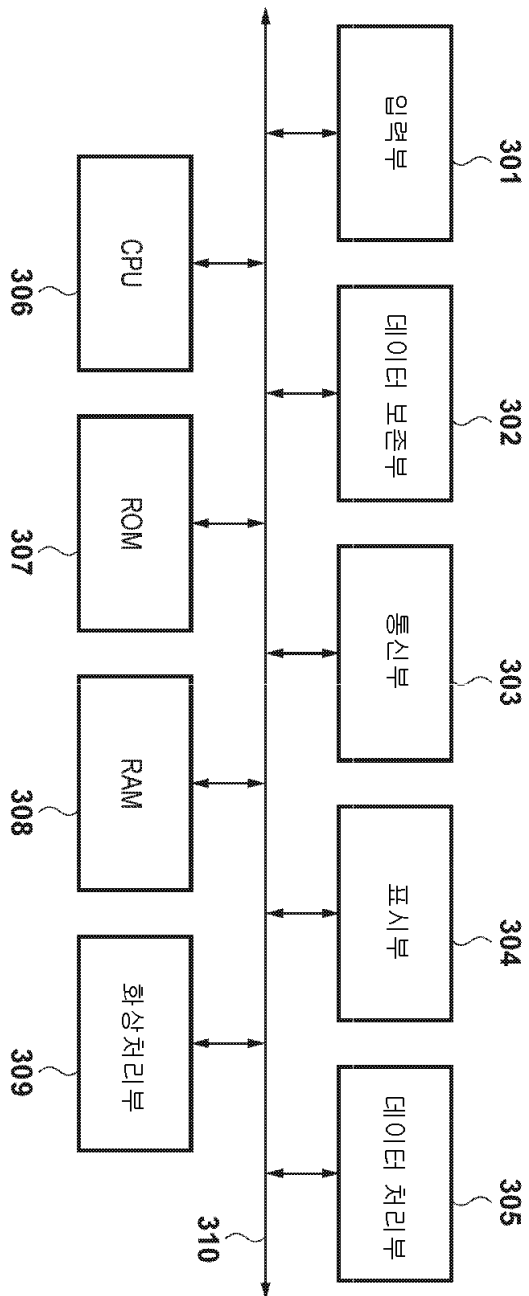
도면1



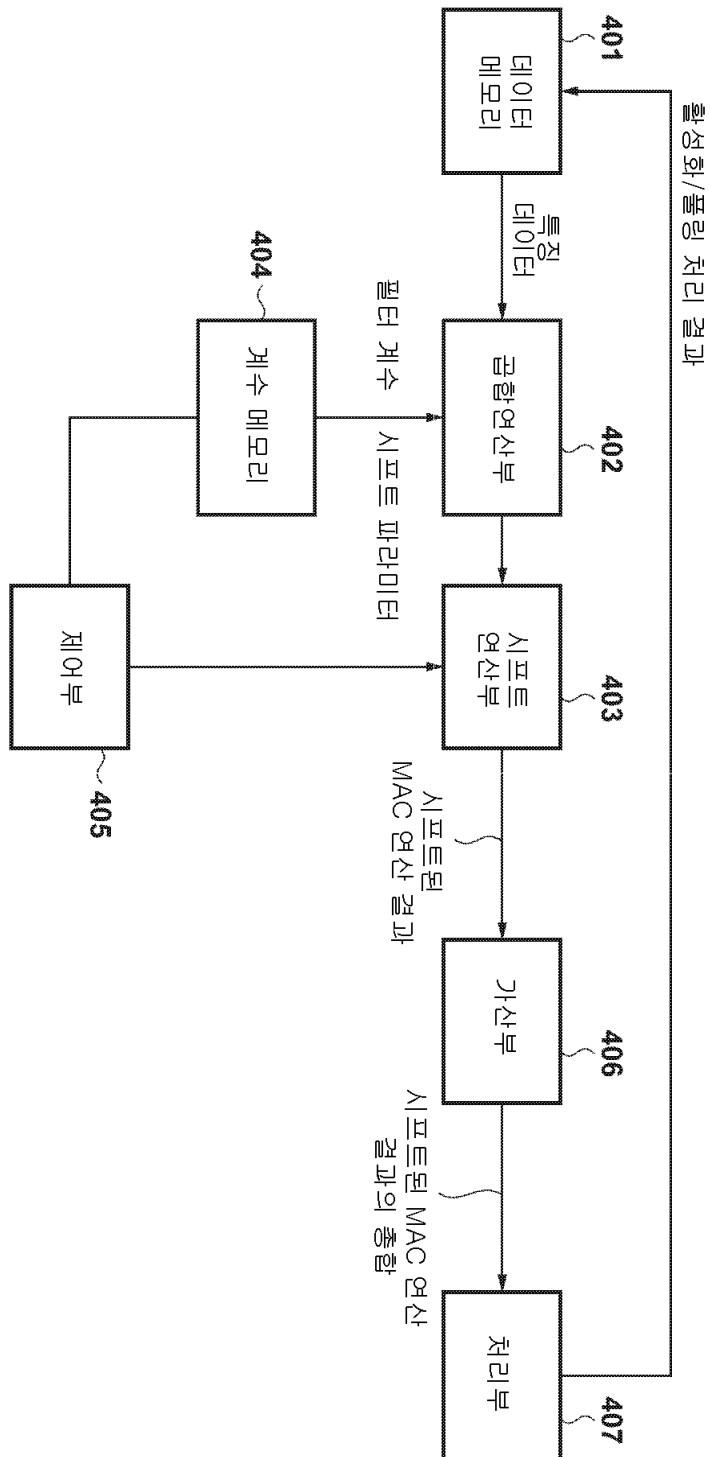
도면2



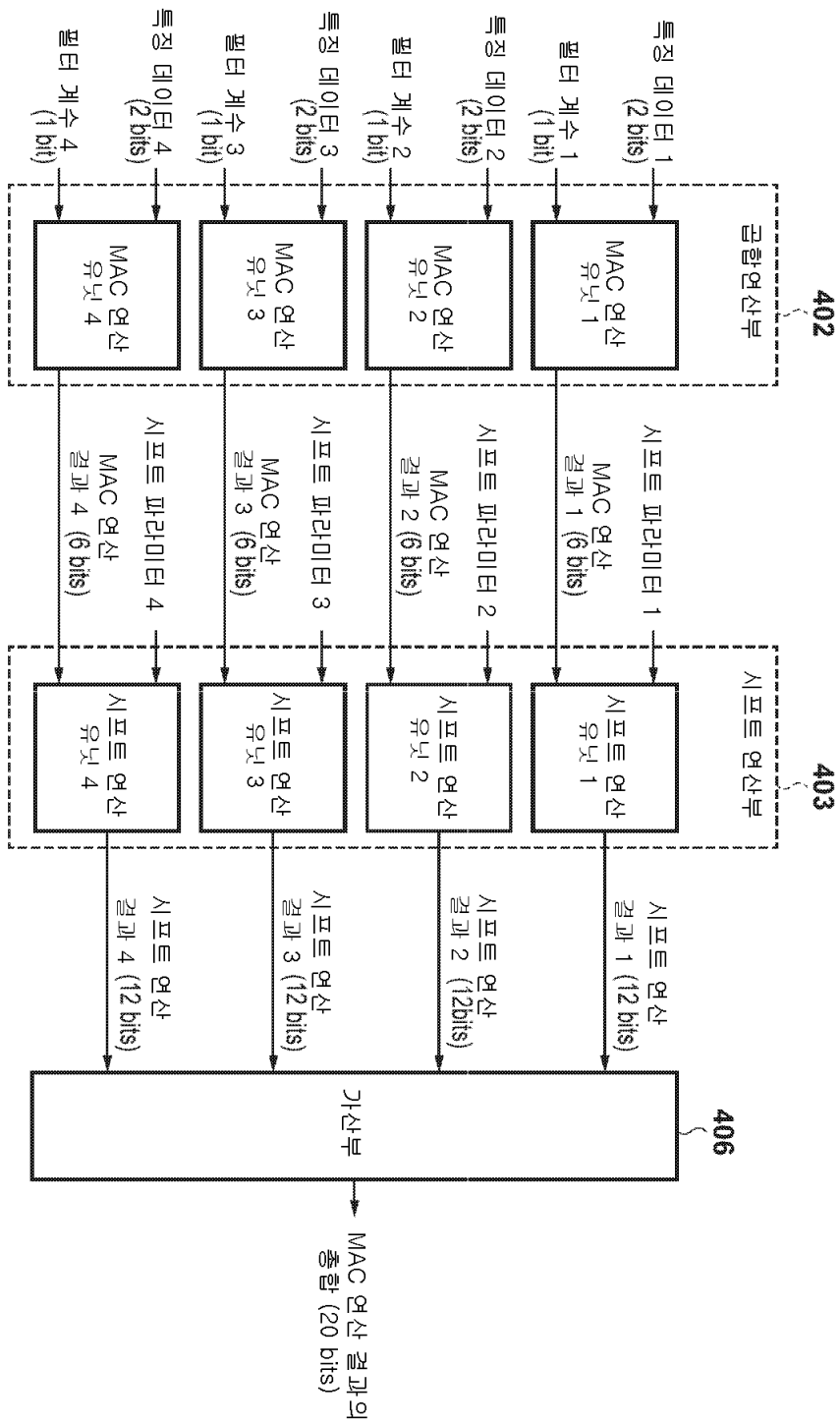
도면3



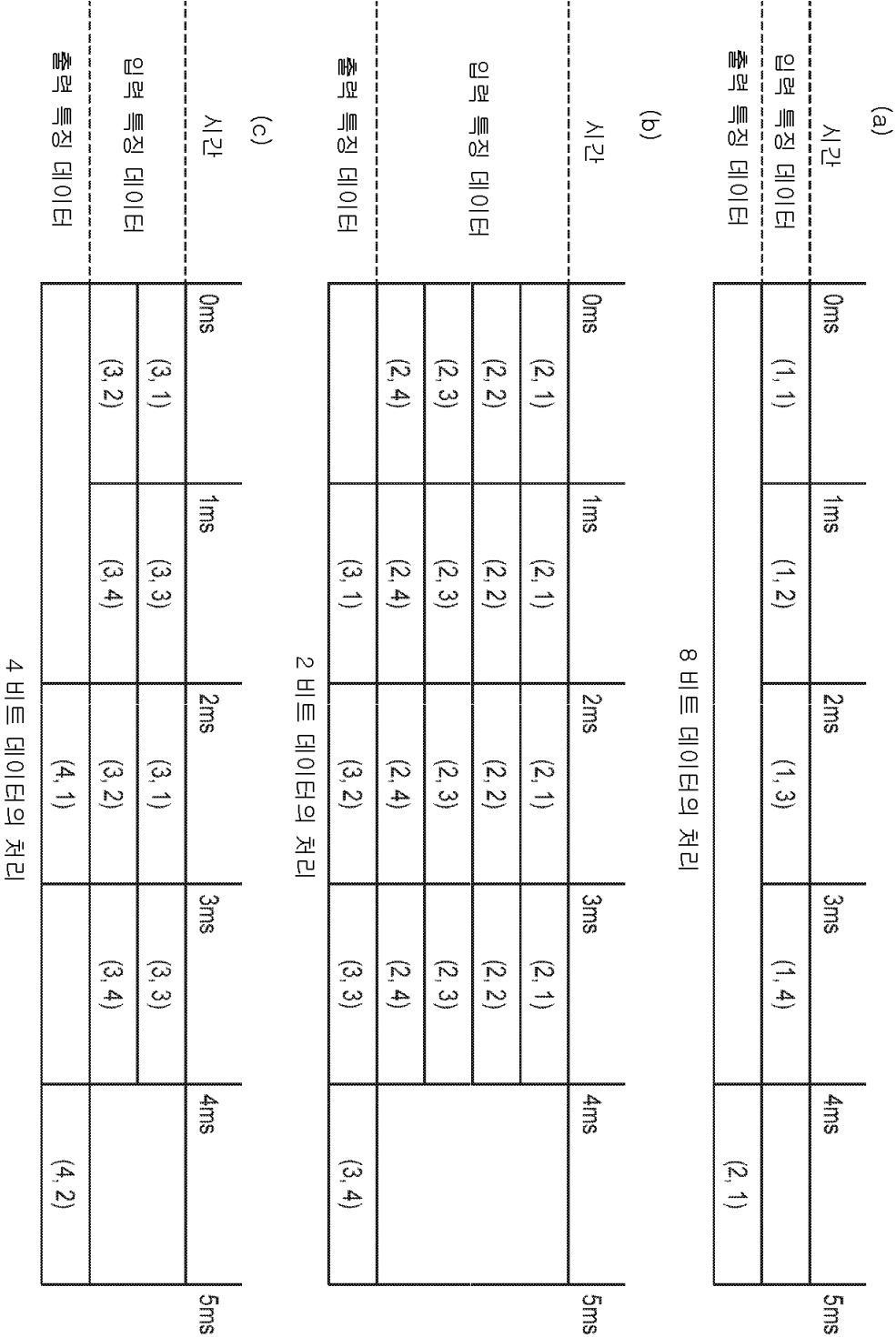
도면4



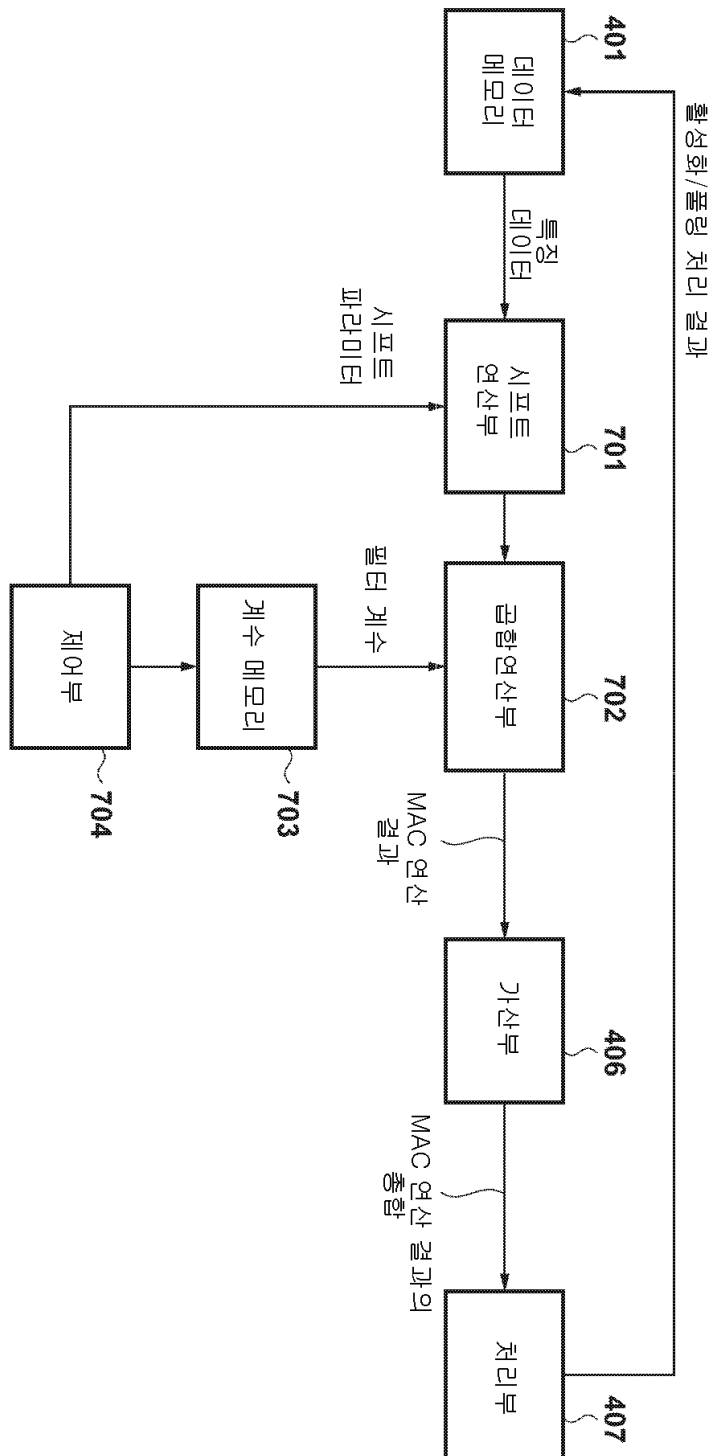
도면5



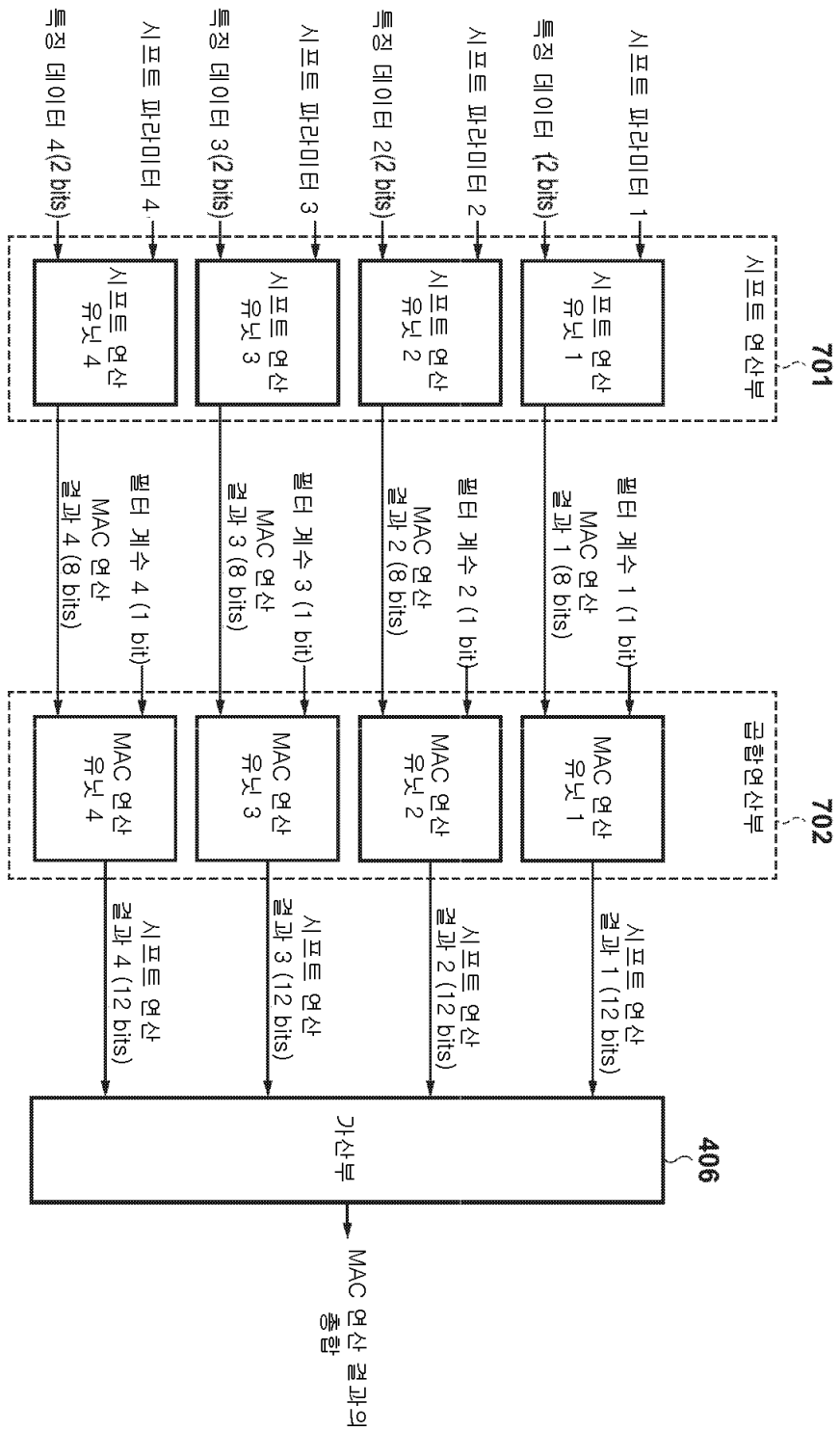
도면6



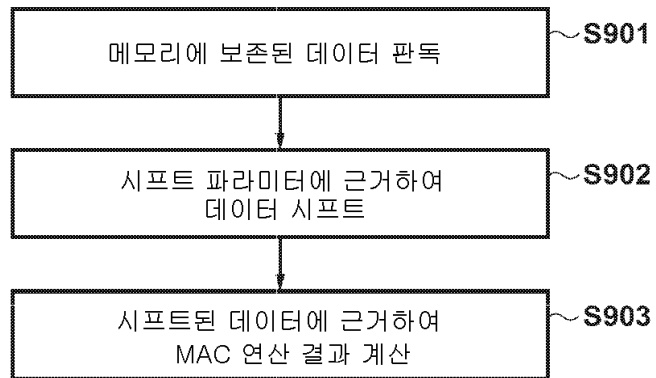
도면7



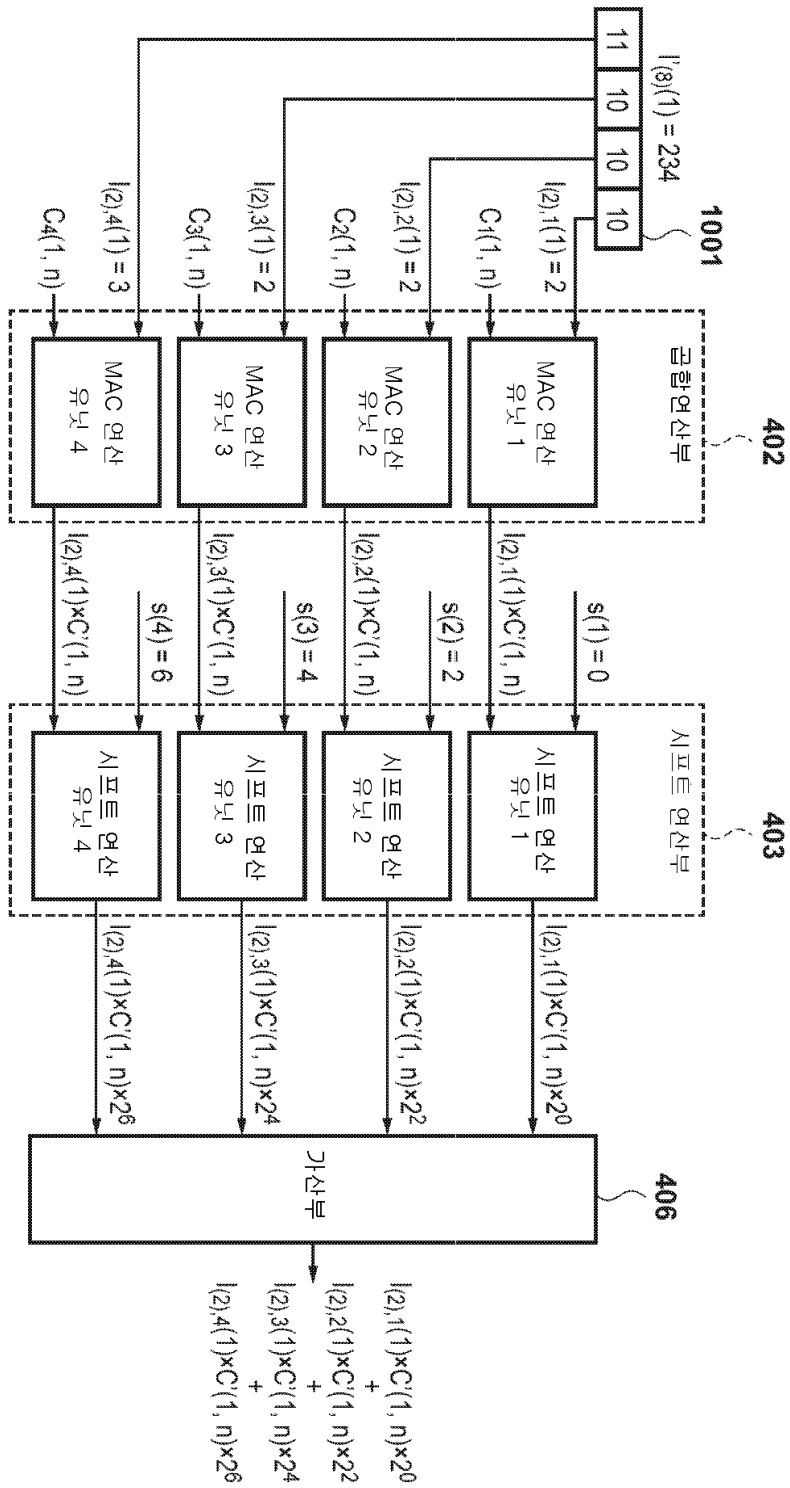
도면8



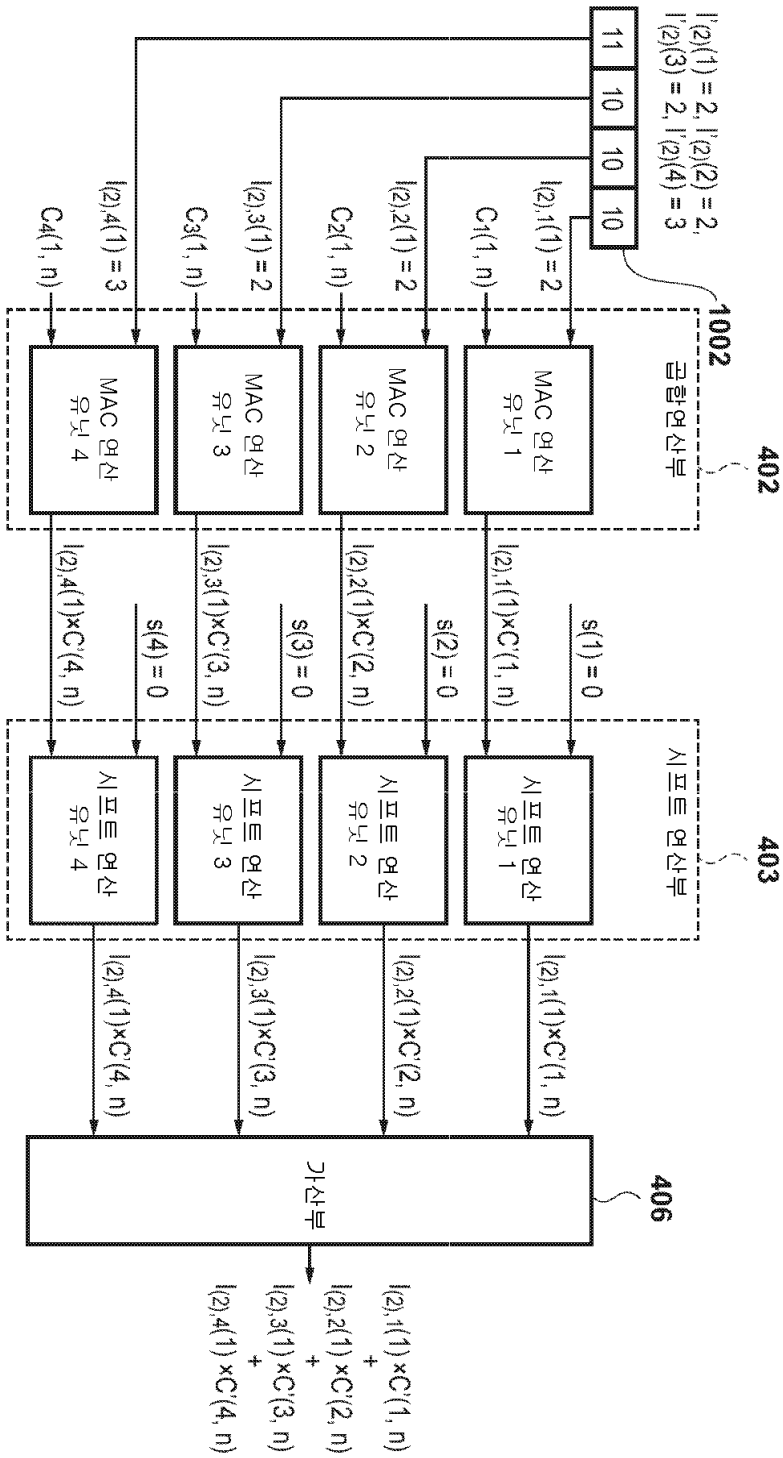
도면9



도면10a



도면10b



도면10c

