



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0127261
(43) 공개일자 2022년09월19일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) G06N 3/04 (2006.01)
G06N 3/08 (2006.01) H04N 19/124 (2014.01)
H04N 19/13 (2014.01) H04N 19/70 (2014.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2022-7025245
(22) 출원일자(국제) 2022년12월21일
심사청구일자 2022년07월20일
(85) 번역문제출일자 2022년07월20일
(86) 국제출원번호 PCT/EP2020/087489
(87) 국제공개번호 WO 2021/123438
국제공개일자 2021년06월24일
(30) 우선권주장
19218862.1 2019년12월20일
유럽특허청(EPO)(EP)

(71) 출원인
프라운호퍼 게젤샤프트 쭈르 뢰르데룽 데어 안겐
반텐 포르슘 에. 베.
독일 80686 뮌헨 한자슈트라쎄 27 체
(72) 발명자
하제, 파울
독일 10587 베를린 아인슈타인우퍼 37 에이치에이
치아이 하인리히-헤르쯔-인스티투트 프라운호퍼-
인스티투트 뢰어 나흐리히텐테크니크 내
키르호호퍼, 하이너
독일 10587 베를린 아인슈타인우퍼 37 에이치에이
치아이 하인리히-헤르쯔-인스티투트 프라운호퍼-
인스티투트 뢰어 나흐리히텐테크니크 내
(뒷면에 계속)
(74) 대리인
윤의섭

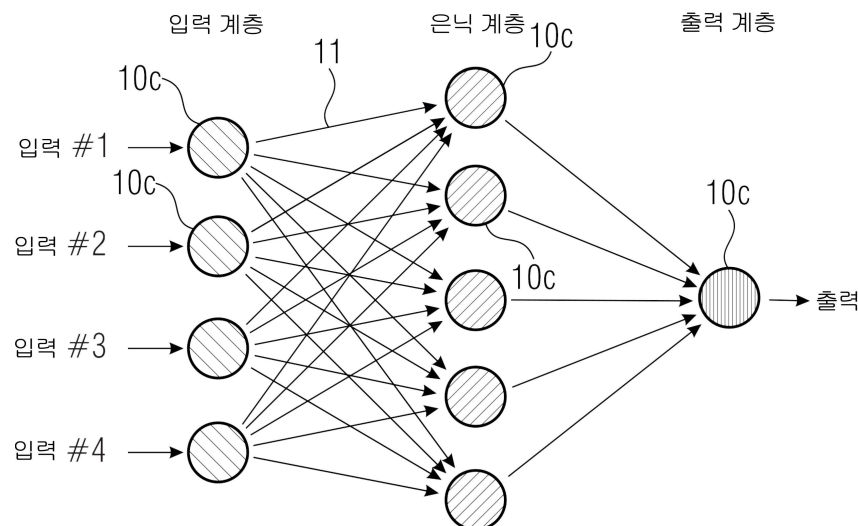
전체 청구항 수 : 총 111 항

(54) 발명의 명칭 신경망 매개변수 코딩에 대한 개념

(57) 요약

본 발명의 제1 측면에 따른 실시예는 신경망 매개변수가 일정하지 않은 양자화기를 사용하지만, 신경망 매개변수를 코딩하는 동안에 이를 변경하여, 즉, 이전 또는 이전에 인코딩된 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩되거나 이에 인코딩된 양자화 인덱스에 따라 복원 레벨 세트를 선택함으로써, 보다 효율적으로 압축될 수 있다는 아이디어에 기초한다. 본 발명의 제2 측면에 따른 실시예는 신경층에서 신경망의 계층화된 구성과 구별하기 위해 복원 계층이라고 불리는 단계별로 수행될 때, 이들 단계에서 제공되는 매개변수화가 신경망 매개변수별로 결합되어 다른 단계에 비해 개선된 신경망 매개변수화를 생성하는 경우 보다 효율적인 신경망 코딩을 달성할 수 있다는 아이디어에 기초한다.

대표도 - 도1



(52) CPC특허분류

G06N 3/084 (2013.01)

H04N 19/124 (2015.01)

H04N 19/13 (2015.01)

H04N 19/70 (2015.01)

(72) 발명자

슈바르츠, 하이코

독일 10587 베를린 아인슈타인우퍼 37 에이치에이
치아이 하인리히-헤르쯔-인스티투트 프라운호퍼-인
스티투트 뢰어 나흐리히텐테크니크 내

마르페, 데틀레프

독일 10587 베를린 아인슈타인우퍼 37 에이치에이
치아이 하인리히-헤르쯔-인스티투트 프라운호퍼-인
스티투트 뢰어 나흐리히텐테크니크 내

비간트, 토마스

독일 10587 베를린 아인슈타인우퍼 37 에이치에이
치아이 하인리히-헤르쯔-인스티투트 프라운호퍼-인
스티투트 뢰어 나흐리히텐테크니크 내

명세서

청구범위

청구항 1

데이터 스트림(14)으로부터 신경망(10)을 정의하는 신경망 매개변수(13)를 디코딩하기 위한 장치에 있어서,
 현재 신경망 매개변수(13')에 대해, 이전 신경망 매개변수에 대해 상기 데이터 스트림(14)으로부터 디코딩된 양자화 인덱스(58)에 따라, 복수(50)의 복원 레벨 세트(52) 중에서 복원 레벨 세트(48)를 선택하고(54),
 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩하고, -
 상기 양자화 인덱스(56)는 상기 현재 신경망 매개변수에 대해 상기 선택된 복원 레벨의 세트(48) 중 하나의 복원 레벨을 나타냄 - ,
 상기 현재 신경망 매개변수(13')를 상기 현재 신경망 매개변수에 대해 상기 양자화 인덱스(56)에 의해 표시되는 상기 선택된 복원 레벨의 세트(48) 중 상기 하나의 복원 레벨로 역양자화함으로써(62),
 상기 신경망 매개변수(13)를 순차적으로 디코딩하도록 구성되는, 장치.

청구항 2

제 1 항에 있어서, 상기 신경망 매개변수(13)는 상기 신경망(10)의 뉴런 상호 연결(11)의 가중치와 관련되는, 장치.

청구항 3

제 1 항 또는 제 2 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2인, 장치.

청구항 4

제 1 항 내지 제 3 항 중 어느 한 항에 있어서, 미리 결정된 양자화 단계 크기(QP)를 통해 상기 복수(50)의 복원 레벨 세트(52)를 매개변수화하고(60) 상기 데이터 스트림(14)으로부터 상기 미리 결정된 양자화 단계 크기에 대한 정보를 유도하도록 구성되는, 장치.

청구항 5

선행 항들 중 어느 한 항에 있어서, 상기 신경망은 하나 이상의 NN 계층을 포함하고 상기 장치는,
 각 NN 계층(p ; $p-1$)에 대해, 상기 각 NN 계층에 대해 미리 결정된 양자화 단계 크기에 대한 정보를 상기 데이터 스트림(14)으로부터 유도하고,
 각각의 NN 계층에 대해, 상기 각각의 NN 계층에 속하는 상기 신경망 매개변수를 역양자화하는 데 사용되기 위해 상기 각각의 NN 계층에 대해 유도된 상기 미리 결정된 양자화 단계 크기를 사용하여 상기 복수(50)의 복원 레벨 세트(52)를 매개변수화하도록
 구성되는, 장치.

청구항 6

제 1 항 내지 제 5 항 중 어느 한 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고 상기 복수의 복원 레벨 세트는,
 미리 결정된 양자화 단계 크기의 0 및 짝수 배수를 포함하는 제1 복원 레벨 세트(세트 0), 및
 미리 결정된 양자화 단계 크기의 0 및 홀수 배수를 포함하는 제2 복원 레벨 세트(세트 1)
 를 포함하는, 장치.

청구항 7

제 1 항 내지 제 6 항 중 어느 한 항에 있어서, 모든 복원 레벨 세트의 모든 복원 레벨은 미리 결정된 양자화 단계 크기의 정수배를 나타내고, 상기 장치는,

각각의 신경망 매개변수에 대해, 상기 각각의 신경망 매개변수에 대해 설정된 상기 선택된 복원 레벨 및 상기 각각의 신경망 매개변수에 대한 상기 엔트로피 디코딩된 양자화 인덱스에 따라 중간 정수 값을 유도하고,

각각의 신경망 매개변수에 대해, 상기 각각의 신경망 매개변수에 대한 상기 중간 값을 상기 각각의 신경망 매개변수에 대한 상기 미리 결정된 양자화 단계 크기와 곱함으로써,

상기 신경망 매개변수를 역양자화하도록 구성되는, 장치.

청구항 8

제 7 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고, 상기 장치는,

상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제 1 세트인 경우, 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스에 2를 곱하여 상기 각각의 신경망 매개변수에 대한 상기 중간 값을 획득하고;

상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제2 세트이고 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스가 0인 경우, 상기 각각의 샘플에 대한 상기 중간 값을 0으로 설정하고;

상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제2 세트이고 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스가 0보다 큰 경우, 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스에 2를 곱하고 상기 곱의 결과에서 1을 빼서 상기 각각의 신경망 매개변수에 대한 상기 중간 값을 획득하고;

현재 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제 2 세트이고 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스가 0보다 작은 경우, 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스에 2를 곱하고 상기 곱의 결과에 1을 더하여 상기 각 신경망 매개변수에 대한 중간 상기 값을 획득함으로써,

각각의 신경망 매개변수에 대한 상기 중간 값을 유도하도록 구성되는, 장치.

청구항 9

제 1 항 내지 제 8 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해, LSB 부분 또는 이전에 디코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로부터 디코딩된 상기 양자화 인덱스(58)의 이전화의 이전에 디코딩된 빈에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 10

제 1 항 내지 제 8 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해, 이전에 디코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로부터 디코딩된 상기 양자화 인덱스(58)의 이전 함수의 결과에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 11

제 1 항 내지 제 10 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해, 이전에 디코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로부터 디코딩된 상기 양자화 인덱스(58)의 패리티에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 12

제 1 항 내지 제 11 항 중 어느 한 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고, 상기 장치는,

상기 각각의 신경망 매개변수에 대한 상기 선택된 복원 레벨 세트 및 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스의 이전 함수에 기초하여 각각의 신경망 매개변수에 대한 하위 세트 인덱스를 유도하여, 결과적으로 상기 하위 세트 인덱스에 대해 4개의 가능한 값이 생성되게 하고,

상기 현재 신경망 매개변수(13')에 대해, 이전에 디코딩된 신경망 매개변수에 대한 상기 하위 세트 인덱스에 따

라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 13

제 12 항에 있어서, 상기 장치는,

상기 현재 신경망 매개변수(13')에 대해, 바로 이전에 디코딩된 여러 신경망 매개변수에 대한 상기 하위 세트 인덱스에 의존하는 선택 규칙을 사용하여 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하고, 상기 현재 신경망 매개변수(13')에 대해, 신경망 매개변수의 전체 또는 일부에 대해 상기 선택 규칙을 사용하도록 구성되는, 장치.

청구항 14

제 13 항에 있어서, 상기 선택 규칙이 의존하는 바로 이전에 디코딩된 신경망 매개변수의 수는 2인, 장치.

청구항 15

제 12 항 내지 제 14 항 중 어느 한 항에 있어서, 각 신경망 매개변수에 대한 상기 하위 세트 인덱스는 상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트 및 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스의 패리티에 기초하여 유도되는, 장치.

청구항 16

제 1 항 내지 제 15 항 중 어느 한 항에 있어서, 상기 장치는,

상기 현재 신경망 매개변수(13')에 대해, 상기 현재 신경망 매개변수(13')와 연관된 상태에 따라 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 결정하고,

바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로부터 디코딩된 상기 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상기 상태를 업데이트함으로써,

상기 현재 신경망 매개변수(13')에 대해, 상태 전이 프로세스를 통해 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 17

제 16 항에 있어서, 상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로부터 디코딩된 상기 양자화 인덱스(58)의 이진 함수를 사용하여 상기 후속 신경망 매개변수에 대한 상기 상태를 업데이트하도록 구성되는, 장치.

청구항 18

제 16 항에 있어서, 상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로부터 디코딩된 상기 양자화 인덱스(58)의 패리티를 사용하여 상기 후속 신경망 매개변수에 대한 상기 상태를 업데이트하도록 구성되는, 장치.

청구항 19

제 16 항 내지 제 18 항 중 어느 한 항에 있어서, 상기 상태 전이 과정은 4개 또는 8개의 가능한 상태 사이에서 전이하도록 구성되는, 장치.

청구항 20

제 16 항 내지 제 19 항 중 어느 한 항에 있어서, 상기 상태 전이 과정에서 짝수 개의 가능한 상태 사이에서 전이하도록 구성되고, 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고, 상기 현재 신경망 매개변수(13')에 대해, 상기 현재 신경망 매개변수(13')와 연관된 상기 상태에 따라 상기 양자화 세트(52) 중에서 양자화 레벨의 세트(48)를 결정하는 단계는,

상기 상태가 짝수개의 가능한 상태 중 전반에 속하는 경우 상기 복수(50)의 복원 레벨 세트(52) 중에서 제1 복

원 레벨 세트를 결정(52)하고, 상기 상태가 상기 짝수개의 가능한 상태 중 후반에 속하는 경우 상기 복수(50)의 복원 레벨 세트(52) 중에서 설정된 제2 복원 레벨을 결정하는, 장치.

청구항 21

제 16 항 내지 제 20 항 중 어느 한 항에 있어서, 상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로부터 디코딩된 상기 양자화 인덱스(58)의 패리티와 상태의 조합을 상기 후속 신경망 매개변수와 관련된 추가 상태에 매핑하는 전이 표를 사용하여 상기 상태의 상기 업데이트를 수행하도록 구성되는, 장치.

청구항 22

제 1 항 내지 제 21 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48)에 좌우하는(123) 확률 모델을 사용하는 산술 코딩 사용하여 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 디코딩하도록 구성되는, 장치.

청구항 23

제 1 항 내지 제 21 항 중 어느 한 항에 있어서,

상기 현재 신경망 매개변수(13')에 대해, 상기 현재 신경망 매개변수(13')와 관련된 상태에 따라 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 결정하고

상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로부터 디코딩된 상기 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상기 상태를 업데이트함으로써,

상기 현재 신경망 매개변수(13')에 대해, 상태 전이 프로세스를 통해 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 선택(54)하고,

상기 현재 신경망 매개변수의 상기 상태(13')에 의존하는(122) 확률 모델을 사용하는 산술 코딩을 사용하여 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 디코딩하도록 구성되는, 장치.

청구항 24

제 23 항에 있어서, 상기 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈(84)에 대한 상기 현재 신경망 매개변수(13')에 대해 상기 상태(122)에 의존하는(122) 상기 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 디코딩하도록 구성되는, 장치.

청구항 25

제 23 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0과 동일하거나 동일하지 않음을 나타내는 유의성 빈을 포함하는, 장치.

청구항 26

제 23 항 내지 제 25 항 중 어느 한 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0보다 크거나 0보다 작음을 나타내는 부호 빈(86)을 포함하는, 장치.

청구항 27

제 23 항 내지 제 26 항 중 어느 한 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)의 절대값이 X보다 크거나 크지 않음을 나타내는, X 초과 빈을 포함하고, 여기서 X는 0보다 큰 정수인, 장치.

청구항 28

제 22 항, 제 23 항 및 제 25 항 내지 27 항 중 어느 한 항에 있어서, 상기 확률 모델의 의존성은 상기 의존성을 사용하여 상기 신경망 매개변수에 대한 컨텍스트 세트 중에서 하나의 컨텍스트(87)의 선택(103)을 포함하도

록 구성되고, 각 컨텍스트는 미리 결정된 확률 모델이 관련되는, 장치.

청구항 29

제 28 항에 있어서, 상기 컨텍스트 각각과 연관된 상기 미리 결정된 확률 모델을 상기 각각의 컨텍스트를 사용하여 산술적으로 코딩된 상기 양자화 인덱스에 기초하여 업데이트하도록 구성되는, 장치.

청구항 30

제 1 항 내지 제 29 항 중 어느 한 항에 있어서, 상기 양자화 인덱스의 이진화의 적어도 하나의 빈에 대해 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48)에 의존하는 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 디코딩하도록 구성되는, 장치.

청구항 31

제 30 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0과 동일하거나 동일하지 않음을 나타내는 유의성 빈을 포함하는, 장치.

청구항 32

제 30 항 또는 제 31 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0보다 크거나 0보다 작음을 나타내는 부호 빈을 포함하는, 장치.

청구항 33

제 30 항 내지 제 32 항 중 어느 한 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)의 절대값이 X보다 크거나 크지 않음을 나타내는, X 초과 빈을 포함하고, 여기서 X는 0보다 큰 정수인, 장치.

청구항 34

제 22 항 내지 제 33 항 중 어느 한 항에 있어서, 상기 확률 모델은 이전에 디코딩된 신경망 매개변수의 상기 양자화 인덱스에 추가로 의존하는, 장치.

청구항 35

제 34 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48) 및 상기 상태에 따라, 복수의 확률 모델 중에서 확률 모델의 하위 세트를 미리 선택하고 이전에 디코딩된 신경망 매개변수의 상기 양자화 인덱스에 의존하여(121) 상기 확률 모델의 하위 세트 중에서 상기 현재 신경망 매개변수에 대한 상기 확률 모델을 선택하도록 구성되는, 장치.

청구항 36

제 35 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48) 또는 상기 상태에 따라, 제1 상태 또는 복원 레벨 세트에 대해 미리 선택된 하위 세트가 임의의 다른 상태 또는 복원 레벨 세트에 대해 미리 선택된 하위 세트와 분리되도록 하는 방식으로 상기 복수의 확률 모델 중에서 상기 확률 모델의 하위 세트를 미리 선택하도록 구성되는, 장치.

청구항 37

제 35 항 또는 제 36 항에 있어서, 상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 상기 양자화 인덱스에 따라 상기 확률 모델의 하위 세트 중에서 상기 현재 신경망 매개변수에 대한 상기 확률 모델을 선택하도록 구성되는, 장치.

청구항 38

제 35 항 또는 제 36 항에 있어서, 상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 상기 양자화 인덱스의 특성에 따라 상기 확률 모델의 하위 세트 중

에서 상기 현재 신경망 매개 변수에 대한 상기 확률 모델을 선택하도록 구성되고, 상기 특성은,

상기 현재 신경망 매개변수가 관련된 부분에 인접한 상기 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개 변수의 0이 아닌 양자화 인덱스의 상기 부호,

상기 현재 신경망 매개변수가 관련된 부분에 인접한 상기 신경망의 부분과 관련되고 0이 아닌, 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 수,

상기 현재 신경망 매개변수가 관련된 부분에 인접한 상기 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개 변수의 양자화 인덱스의 절대값의 합,

상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개 변수의 양자화 인덱스의 절대값의 합과

상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련되고 0이 아닌, 상기 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 수 간의

차이

중 하나 이상을 포함하는, 장치.

청구항 39

제 37 항 또는 제 38 항에 있어서, 상기 이전에 디코딩된 신경망 매개변수가 상기 현재 신경망 매개변수와 동일한 신경망 계층과 관련되도록 상기 이전에 디코딩된 신경망 매개변수를 위치 선정하도록 구성되는, 장치.

청구항 40

제 37 항 또는 제 39 항에 있어서, 상기 하나 이상의 이전에 디코딩된 신경망 매개변수가 상기 현재 신경망 매개변수가 참조하는 뉴런 상호 연결과 관련된 뉴런, 또는 상기 뉴런에 인접한 추가 뉴런에서 나오거나 이것으로 이어지는 뉴런 상호 연결과 관련되도록 하는 방식으로 상기 이전에 디코딩된 신경망 매개변수 중 하나 이상을 위치 선정하도록 구성되는, 장치.

청구항 41

제 1 항 내지 제 40 항 중 어느 한 항에 있어서, 상기 신경망 매개변수(13)에 대한 상기 양자화 인덱스(56)를 디코딩하고 상기 신경망 매개변수(13) 간의 공통 순차적 순서(14')를 따라 상기 신경망 매개변수(13)의 역양자화를 수행하도록 구성되는, 장치.

청구항 42

제 1 항 내지 제 41 항 중 어느 한 항에 있어서, 상기 양자화 인덱스의 이진화의 하나 이상의 선행 빈에 대해 이전에 디코딩된 신경망 매개변수에 의존하는 상기 확률 모델을 사용함으로써 및 상기 하나 이상의 선행 빈을 따르는 상기 양자화 인덱스의 상기 이진화의 동등 확률 바이패스 모드 접미사 빈을 사용함으로써 이진 산술 코딩을 사용하여 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 디코딩하도록 구성되는, 장치.

청구항 43

제 42 항에 있어서, 상기 양자화 인덱스의 상기 이진화의 상기 접미사 빈은, 그 절대 값이 상기 하나 이상의 선행 빈으로 표현할 수 있는 최대 절대값을 초과하는 상기 양자화 인덱스의 값을 이진화하기 위한 접미사 이진화의 이진화 코드의 빈을 나타내고, 상기 장치는 이전에 디코딩된 신경망 매개변수의 상기 양자화 인덱스에 따라 상기 상기 접미사 이진화를 선택하도록 구성되는, 장치.

청구항 44

제 1 항 내지 제 43 항 중 어느 한 항에 있어서, 상기 신경망 매개변수는 상기 신경망(10)이 표현되는 복원 계층의 하나의 복원 계층에 관한 것이고, 상기 장치는,

상기 신경망 매개변수 별로 상기 신경망 매개변수를 하나 이상의 추가 복원 계층의 대응하는 신경망 매개변수와

결합하여 상기 신경망을 복원하도록 구성되는, 장치.

청구항 45

제 44 항에 있어서, 상기 현재 신경망 매개변수에 해당하는 해당 신경망 매개변수에 의존하는 확률 모델을 사용하는 산술 코딩을 사용하여 상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 디코딩하도록 구성되는, 장치.

청구항 46

데이터 스트림(14)으로부터 신경망(10)을 정의하는 신경망 매개변수(13)를 인코딩하기 위한 장치에 있어서, 현재 신경망 매개변수(13')에 대해, 이전에 인코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로 인코딩된 양자화 인덱스(58)에 따라, 복수(50)의 복원 레벨 세트(52) 중에서 복원 레벨 세트(48)를 선택하고(54), 상기 현재 신경망 매개변수(13')를 상기 선택된 복원 레벨 세트(48) 중 상기 하나의 복원 레벨로 양자화하고(64),

상기 현재 신경망 매개변수에 대한 양자화 인덱스(56)가 상기 데이터 스트림(14)으로 양자화되는 상기 하나의 복원 레벨을 나타내는 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 인코딩함으로써,

상기 신경망 매개변수(13)를 순차적으로 인코딩하도록 구성되는, 장치.

청구항 47

제 46 항에 있어서, 상기 신경망 매개변수(13)는 상기 신경망(10)의 뉴런 상호 연결(11)의 가중치와 관련되는, 장치.

청구항 48

제 46 항 또는 제 47 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2인, 장치.

청구항 49

제 46 항 내지 제 48 항 중 어느 한 항에 있어서, 미리 결정된 양자화 단계 크기(QP)를 통해 상기 복수(50)의 복원 레벨 세트(52)를 매개변수화하고(60) 상기 데이터 스트림(14)으로 상기 미리 결정된 양자화 단계 크기에 대한 정보를 삽입하도록 구성되는, 장치.

청구항 50

선행 항들 중 어느 한 항에 있어서, 상기 신경망은 하나 이상의 NN 계층을 포함하고 상기 장치는,

각 NN 계층(p ; $p-1$)에 대해, 상기 각 NN 계층에 대해 미리 결정된 양자화 단계 크기에 대한 정보를 상기 데이터 스트림(14)으로 삽입하고,

각각의 NN 계층에 대해, 상기 각각의 NN 계층에 속하는 상기 신경망 매개변수를 양자화하는 데 사용되기 위해 상기 각각의 NN 계층에 대해 유도된 상기 미리 결정된 양자화 단계 크기를 사용하여 상기 복수(50)의 복원 레벨 세트(52)를 매개변수화하도록

구성되는, 장치.

청구항 51

제 46 항 내지 제 50 항 중 어느 한 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고 상기 복수의 복원 레벨 세트는,

미리 결정된 양자화 단계 크기의 0 및 짝수 배수를 포함하는 제1 복원 레벨 세트(세트 0), 및

미리 결정된 양자화 단계 크기의 0 및 홀수 배수를 포함하는 제2 복원 레벨 세트(세트 1)

를 포함하는, 장치.

청구항 52

제 46 항 내지 제 51 항 중 어느 한 항에 있어서, 모든 복원 레벨 세트의 모든 복원 레벨은 미리 결정된 양자화 단계 크기의 정수배를 나타내고, 상기 장치는,

각각의 신경망 매개변수에 대해, 상기 각각의 신경망 매개변수에 대해 설정된 상기 선택된 복원 레벨 및 상기 각각의 신경망 매개변수에 대한 상기 엔트로피 인코딩된 양자화 인덱스에 따라 중간 정수 값을 유도하고,

각각의 신경망 매개변수에 대해, 상기 각각의 신경망 매개변수에 대한 상기 중간 값을 상기 각각의 신경망 매개변수에 대한 상기 미리 결정된 양자화 단계 크기와 곱함으로써,

역양자화하도록 하는 방식으로 상기 신경망 매개변수를 양자화하도록 구성되는, 장치.

청구항 53

제 52 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고, 상기 장치는,

상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제 1 세트인 경우, 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스에 2를 곱하여 상기 각각의 신경망 매개변수에 대한 상기 중간 값을 획득하고;

상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제2 세트이고 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스가 0인 경우, 상기 각각의 샘플에 대한 상기 중간 값을 0으로 설정하고;

상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제2 세트이고 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스가 0보다 큰 경우, 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스에 2를 곱하고 상기 곱의 결과에서 1을 빼서 상기 각각의 신경망 매개변수에 대한 상기 중간 값을 획득하고;

현재 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트가 제 2 세트이고 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스가 0보다 작은 경우, 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스에 2를 곱하고 상기 곱의 결과에 1을 더하여 상기 각 신경망 매개변수에 대한 중간 상기 값을 획득함으로써,

각각의 신경망 매개변수에 대한 상기 중간 값을 유도하도록 구성되는, 장치.

청구항 54

제 46 항 내지 제 53 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해, LSB 부분 또는 이전에 인코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로 인코딩된 상기 양자화 인덱스(58)의 이전화의 이전에 인코딩된 빈에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 55

제 46 항 내지 제 53 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해, 이전에 인코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로 인코딩된 상기 양자화 인덱스(58)의 이전 함수의 결과에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 56

제 46 항 내지 제 55 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해, 이전에 인코딩된 신경망 매개변수에 대해 상기 데이터 스트림(14)으로 인코딩된 상기 양자화 인덱스(58)의 패리티에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 57

제 46 항 내지 제 56 항 중 어느 한 항에 있어서, 상기 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고, 상기 장치는,

상기 각각의 신경망 매개변수에 대한 상기 선택된 복원 레벨 세트 및 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스의 이전 함수에 기초하여 각각의 신경망 매개변수에 대한 하위 세트 인덱스를 유도하여, 결과적으

로 상기 하위 세트 인덱스에 대해 4개의 가능한 값이 생성되게 하고,

상기 현재 신경망 매개변수(13')에 대해, 이전에 인코딩된 신경망 매개변수에 대한 상기 하위 세트 인덱스에 따라, 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하도록

구성되는, 장치.

청구항 58

제 57 항에 있어서, 상기 장치는,

상기 현재 신경망 매개변수(13')에 대해, 바로 이전에 인코딩된 여러 신경망 매개변수에 대한 상기 하위 세트 인덱스에 의존하는 선택 규칙을 사용하여 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 복원 레벨 세트(48)를 선택(54)하고, 상기 현재 신경망 매개변수(13')에 대해, 신경망 매개변수의 전체 또는 일부에 대해 상기 선택 규칙을 사용하도록 구성되는, 장치.

청구항 59

제 58 항에 있어서, 상기 선택 규칙이 의존하는 바로 이전에 인코딩된 신경망 매개변수의 수는 2인, 장치.

청구항 60

제 57 항 내지 제 59 항 중 어느 한 항에 있어서, 각 신경망 매개변수에 대한 상기 하위 세트 인덱스는 상기 각각의 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트 및 상기 각각의 신경망 매개변수에 대한 상기 양자화 인덱스의 패리티에 기초하여 유도되는, 장치.

청구항 61

제 46 항 내지 제 60 항 중 어느 한 항에 있어서, 상기 장치는,

상기 현재 신경망 매개변수(13')에 대해, 상기 현재 신경망 매개변수(13')와 연관된 상태에 따라 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 결정하고,

바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로 인코딩된 상기 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상기 상태를 업데이트함으로써,

상기 현재 신경망 매개변수(13')에 대해, 상태 전이 프로세스를 통해 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 선택(54)하도록 구성되는, 장치.

청구항 62

제 61 항에 있어서, 상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로 인코딩된 상기 양자화 인덱스(58)의 이전 함수를 사용하여 상기 후속 신경망 매개변수에 대한 상기 상태를 업데이트하도록 구성되는, 장치.

청구항 63

제 61 항에 있어서, 상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로 인코딩된 상기 양자화 인덱스(58)의 패리티를 사용하여 상기 후속 신경망 매개변수에 대한 상기 상태를 업데이트하도록 구성되는, 장치.

청구항 64

제 61 항 내지 제 63 항 중 어느 한 항에 있어서, 상기 상태 전이 과정은 4개 또는 8개의 가능한 상태 사이에서 전이하도록 구성되는, 장치.

청구항 65

제 61 항 내지 제 64 항 중 어느 한 항에 있어서, 상기 상태 전이 과정에서 짝수 개의 가능한 상태 사이에서 전이하도록 구성되고, 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2이고, 상기 현재 신경망 매개변수(13')에 대해, 상기 현재 신경망 매개변수(13')와 연관된 상기 상태에 따라 상기 양자화 세트(52) 중에서

양자화 레벨의 세트(48)를 결정하는 단계는,

상기 상태가 짝수개의 가능한 상태 중 전반에 속하는 경우 상기 복수(50)의 복원 레벨 세트(52) 중에서 제1 복원 레벨 세트를 결정(52)하고, 상기 상태가 상기 짝수개의 가능한 상태 중 후반에 속하는 경우 상기 복수(50)의 복원 레벨 세트(52) 중에서 설정된 제2 복원 레벨을 결정하는, 장치.

청구항 66

제 61 항 내지 제 65 항 중 어느 한 항에 있어서, 상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로 인코딩된 상기 양자화 인덱스(58)의 패리티와 상태의 조합을 상기 후속 신경망 매개변수와 관련된 추가 상태에 매핑하는 전이 표를 사용하여 상기 상태의 상기 업데이트를 수행하도록 구성되는, 장치.

청구항 67

제 46항 내지 제 66 항 중 어느 한 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48)에 좌우하는(123) 확률 모델을 사용하는 산술 코딩 사용하여 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 상기 데이터 스트림(14)으로 인코딩하도록 구성되는, 장치.

청구항 68

제 46 항 내지 제 66 항 중 어느 한 항에 있어서,

상기 현재 신경망 매개변수(13')에 대해, 상기 현재 신경망 매개변수(13')와 관련된 상태에 따라 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 결정하고

상기 바로 이전의 신경망 매개변수에 대해 상기 데이터 스트림으로 인코딩된 상기 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상기 상태를 업데이트함으로써,

상기 현재 신경망 매개변수(13')에 대해, 상태 전이 프로세스를 통해 상기 복수(50)의 복원 레벨 세트(52) 중에서 상기 양자화 레벨 세트(48)를 선택(54)하고,

상기 현재 신경망 매개변수의 상기 상태(13')에 의존하는(122) 확률 모델을 사용하는 산술 코딩을 사용하여 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 상기 데이터 스트림(14)으로 인코딩하도록

구성되는, 장치.

청구항 69

제 68 항에 있어서, 상기 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈(84)에 대한 상기 현재 신경망 매개변수(13')에 대해 상기 상태(122)에 의존하는(122) 상기 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 상기 데이터 스트림(14)으로 인코딩하도록 구성되는, 장치.

청구항 70

제 68 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0과 동일하거나 동일하지 않음을 나타내는 유의성 빈을 포함하는, 장치.

청구항 71

제 68 항 내지 제 70 항 중 어느 한 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0보다 크거나 0보다 작음을 나타내는 부호 빈(86)을 포함하는, 장치.

청구항 72

제 68 항 내지 제 71 항 중 어느 한 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)의 절대값이 X보다 크거나 크지 않음을 나타내는, X 초과 빈을 포함하고, 여기서 X는 0보다 큰 정수인, 장치.

청구항 73

제 67 항, 제 68 항 및 제 70 항 내지 72 항 중 어느 한 항에 있어서, 상기 확률 모델의 의존성은 상기 의존성을 사용하여 상기 신경망 매개변수에 대한 컨텍스트 세트 중에서 하나의 컨텍스트(87)의 선택(103)을 포함하도록 구성되고, 각 컨텍스트는 미리 결정된 확률 모델이 관련되는, 장치.

청구항 74

제 73 항에 있어서, 상기 컨텍스트 각각과 연관된 상기 미리 결정된 확률 모델을 상기 각각의 컨텍스트를 사용하여 산술적으로 코딩된 상기 양자화 인덱스에 기초하여 업데이트하도록 구성되는, 장치.

청구항 75

제 46 항 내지 제 74 항 중 어느 한 항에 있어서, 상기 양자화 인덱스의 이진화의 적어도 하나의 빈에 대해 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48)에 의존하는 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 상기 데이터 스트림(14)으로 인코딩하도록 구성되는, 장치.

청구항 76

제 75 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0과 동일하거나 동일하지 않음을 나타내는 유의성 빈을 포함하는, 장치.

청구항 77

제 75 항 또는 제 76 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)가 0보다 크거나 0보다 작음을 나타내는 부호 빈을 포함하는, 장치.

청구항 78

제 75 항 내지 제 77 항 중 어느 한 항에 있어서, 상기 적어도 하나의 빈은 상기 현재 신경망 매개변수의 상기 양자화 인덱스(56)의 절대값이 X보다 크거나 크지 않음을 나타내는, X 초과 빈을 포함하고, 여기서 X는 0보다 큰 정수인, 장치.

청구항 79

제 67 항 내지 제 78 항 중 어느 한 항에 있어서, 상기 확률 모델은 이전에 인코딩된 신경망 매개변수의 상기 양자화 인덱스에 추가로 의존하는, 장치.

청구항 80

제 79 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48) 및 상기 상태에 따라, 복수의 확률 모델 중에서 확률 모델의 하위 세트를 미리 선택하고 이전에 인코딩된 신경망 매개변수의 상기 양자화 인덱스에 의존하여(121) 상기 확률 모델의 하위 세트 중에서 상기 현재 신경망 매개변수에 대한 상기 확률 모델을 선택하도록 구성되는, 장치.

청구항 81

제 80 항에 있어서, 상기 현재 신경망 매개변수(13')에 대해 선택된 상기 복원 레벨 세트(48) 또는 상기 상태에 따라, 제1 상태 또는 복원 레벨 세트에 대해 미리 선택된 하위 세트가 임의의 다른 상태 또는 복원 레벨 세트에 대해 미리 선택된 하위 세트와 분리되도록 하는 방식으로 상기 복수의 확률 모델 중에서 상기 확률 모델의 하위 세트를 미리 선택하도록 구성되는, 장치.

청구항 82

제 80 항 또는 제 81 항에 있어서, 상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 상기 양자화 인덱스에 따라 상기 확률 모델의 하위 세트 중에서 상기 현재 신경망 매개변수에 대한 상기 확률 모델을 선택하도록 구성되는, 장치.

청구항 83

제 80 항 또는 제 81 항에 있어서, 상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 상기 양자화 인덱스의 특성에 따라 상기 확률 모델의 하위 세트 중에서 상기 현재 신경망 매개 변수에 대한 상기 확률 모델을 선택하도록 구성되고, 상기 특성은,

상기 현재 신경망 매개변수가 관련된 부분에 인접한 상기 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 0이 아닌 양자화 인덱스의 상기 부호,

상기 현재 신경망 매개변수가 관련된 부분에 인접한 상기 신경망의 부분과 관련되고 0이 아닌, 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 수,

상기 현재 신경망 매개변수가 관련된 부분에 인접한 상기 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 절대값의 합,

상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 절대값의 합과

상기 현재 신경망 매개변수가 관련된 부분에 인접하는 상기 신경망의 부분과 관련되고 0이 아닌, 상기 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 수 간의

차이

중 하나 이상을 포함하는, 장치.

청구항 84

제 82 항 또는 제 83 항에 있어서, 상기 이전에 인코딩된 신경망 매개변수가 상기 현재 신경망 매개변수와 동일한 신경망 계층과 관련되도록 상기 이전에 인코딩된 신경망 매개변수를 위치 선정하도록 구성되는, 장치.

청구항 85

제 82 항 또는 제 84 항에 있어서, 상기 하나 이상의 이전에 인코딩된 신경망 매개변수가 상기 현재 신경망 매개변수가 참조하는 뉴런 상호 연결과 관련된 뉴런, 또는 상기 뉴런에 인접한 추가 뉴런에서 나오거나 이것으로 이어지는 뉴런 상호 연결과 관련되도록 하는 방식으로 상기 이전에 인코딩된 신경망 매개변수 중 하나 이상을 위치 선정하도록 구성되는, 장치.

청구항 86

제 46 항 내지 제 75 항 중 어느 한 항에 있어서, 상기 신경망 매개변수(13)에 대한 상기 양자화 인덱스(56)를 인코딩하고 상기 신경망 매개변수(13) 간의 공통 순차적 순서(14')를 따라 상기 신경망 매개변수(13)의 양자화를 수행하도록 구성되는, 장치.

청구항 87

제 46 항 내지 제 86 항 중 어느 한 항에 있어서, 상기 양자화 인덱스의 이진화의 하나 이상의 선행 빈에 대해 이전에 인코딩된 신경망 매개변수에 의존하는 상기 확률 모델을 사용함으로써 및 상기 하나 이상의 선행 빈을 따르는 상기 양자화 인덱스의 상기 이진화의 동등 확률 바이패스 모드 점미사 빈을 사용함으로써 이진 산술 코딩을 사용하여 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 상기 데이터 스트림(14)으로 인코딩하도록 구성되는, 장치.

청구항 88

제 87 항에 있어서, 상기 양자화 인덱스의 상기 이진화의 상기 점미사 빈은, 그 절대 값이 상기 하나 이상의 선행 빈으로 표현할 수 있는 최대 절대값을 초과하는 상기 양자화 인덱스의 값을 이진화하기 위한 점미사 이진화의 이진화 코드의 빈을 나타내고, 상기 장치는 이전에 인코딩된 신경망 매개변수의 상기 양자화 인덱스에 따라 상기 점미사 이진화를 선택하도록 구성되는, 장치.

청구항 89

제 46 항 내지 제 88 항 중 어느 한 항에 있어서, 상기 신경망 매개변수는 상기 신경망(10)이 표현되는 복원 계층의 하나의 복원 계층에 관한 것이고, 상기 장치는,

상기 신경망이 상기 신경망 매개변수 별로 상기 신경망 매개변수를 하나 이상의 추가 복원 계층의 대응하는 신경망 매개변수와 결합하여 복원 가능하도록 구성되는, 장치.

청구항 90

제 89 항에 있어서, 상기 현재 신경망 매개변수에 해당하는 해당 신경망 매개변수에 의존하는 확률 모델을 사용하는 산술 코딩을 사용하여 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 상기 데이터 스트림(14)으로 인코딩하도록 구성되는, 장치.

청구항 91

제 46 항 내지 제 90 항 중 어느 한 항에 있어서, 상기 선택 및/또는 상기 양자화를 수행하기 위해 비터비 알고리즘 및 울-왜곡 비용 측정을 사용하도록 구성되는, 장치.

청구항 92

신경망(10)을 정의하는 신경망 매개변수(13)를 복원하기 위한 장치(310)에 있어서,

제1 복원 계층에 대한 제1 신경망 매개변수를 유도하여 신경망 매개변수(13)당 제1 복원 계층 신경망 매개변수 값을 산출하고,

데이터 스트림(14)으로부터 제2 복원 계층에 대한 제2 신경망 매개변수(13)를 디코딩하여(312) 신경망 매개변수(13)당 제2 복원 계층 신경망 매개변수 값을 산출하고,

각각의 신경망 매개변수에 대해 상기 제1 복원 계층 신경망 매개변수 값과 상기 제2 복원 계층 신경망 매개변수 값을 결합함으로써 상기 신경망 매개변수(13)를 복원(314)하도록

구성되는, 장치.

청구항 93

제 92 항에 있어서,

상기 데이터 스트림 또는 별도의 데이터 스트림으로부터 상기 제1 복원 계층에 대한 상기 제1 신경망 매개변수(13)를 디코딩(316)하고,

상기 제1 및 제2 복원 계층에 대한 개별의 확률 컨텍스트를 사용하는 컨텍스트 적응 엔트로피 디코딩에 의해 상기 데이터 스트림으로부터 상기 제2 복원 계층에 대한 상기 제2 신경망 매개변수(13)를 디코딩하도록

구성되는, 장치.

청구항 94

제 92 항 또는 제 93 항에 있어서,

신경망 매개변수당, 상기 제 1 복원 계층 신경망 매개변수 값 및 상기 제 2 복원 계층 신경망 매개변수 값의 매개변수 값의 매개변수 별 합 또는 매개변수 별 곱에 의해 상기 신경망 매개변수(13)를 복원하도록 구성되는, 장치.

청구항 95

제 92 항 내지 제 94 항 중 어느 한 항에 있어서,

상기 제 1 복원 계층 신경망 매개변수 값에 의존하는 확률 모델을 사용하는 컨텍스트 적응 엔트로피 디코딩에 의해 상기 데이터 스트림으로부터 상기 제 2 복원 계층 신경망 매개변수 값을 디코딩하도록 구성되는, 장치.

청구항 96

제 92 항 내지 제 95 항 중 어느 한 항에 있어서,

컨텍스트 적응 엔트로피 디코딩하고,

상기 제 1 복원 계층 신경망 매개변수 값에 따라 확률 컨텍스트 세트의 집합 중에서 확률 컨텍스트 세트를 선택

하고,

상기 제1 복원 계층 신경망 매개변수 값에 따라 상기 선택된 확률 컨텍스트 세트 중에서 사용될 확률 컨텍스트를 선택함으로써,

상기 데이터 스트림으로부터 상기 제 2 복원 계층 신경망 매개변수 값을 디코딩하도록 구성되는, 장치.

청구항 97

제 96 항에 있어서, 상기 확률 컨텍스트 세트의 집합은 3개의 확률 컨텍스트 세트를 포함하고, 상기 장치는,

상기 제 1 복원 계층 신경망 매개변수 값이 음수이면 상기 확률 컨텍스트 세트의 집합 중에서 제 1 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고,

상기 제 1 복원 계층 신경망 매개변수 값이 양수이면 상기 확률 컨텍스트 세트의 집합 중에서 제 2 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고,

상기 제 1 복원 계층 신경망 매개변수 값이 0인 경우 상기 확률 컨텍스트 세트의 집합 중에서 제 3 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하도록

구성되는, 장치.

청구항 98

제 96 항에 있어서, 상기 확률 컨텍스트 세트의 집합은 2개의 확률 컨텍스트 세트를 포함하고, 상기 장치는,

상기 제1 복원 계층 신경망 매개변수 값이 미리 결정된 값보다 큰 경우 상기 확률 컨텍스트 세트의 집합 중에서 제1 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고, 상기 제1 복원 계층 신경망 매개변수 값이 상기 미리 결정된 값보다 크지 않은 경우, 상기 확률 컨텍스트 세트의 집합 중에서 제2 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하거나,

상기 제1 복원 계층 신경망 매개변수 값의 절대값이 상기 미리 결정된 값보다 큰 경우, 상기 확률 컨텍스트 세트의 집합 중에서 제1 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고, 상기 제1 복원 계층 신경망 매개변수 값의 절대값이 상기 미리 결정된 값보다 크지 않은 경우 상기 확률 컨텍스트 세트의 집합 중에서 제2 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하도록

구성되는, 장치.

청구항 99

신경망 매개변수(13) 당, 제 1 복원 계층 신경망 매개변수 값을 포함하는 제1 복원 계층에 대해 제1 신경망 매개변수(13)를 사용함으로써, 신경망(10)을 정의하는 신경망 매개변수(13)를 인코딩하기 위한 장치(320)에 있어서, 상기 장치는,

제2 복원 계층에 대해, 제2 신경망 매개변수(13)를 신경망 매개변수(13) 당 제2 복원 계층 신경망 매개변수 값을 포함하는 데이터 스트림(14)으로 인코딩(322)하도록 구성되고,

상기 신경망 매개변수(13)는, 각각의 신경망 매개변수에 대해 상기 제1 복원 계층 신경망 매개변수 값과 상기 제2 복원 계층 신경망 매개변수 값을 결합함으로써, 복원 가능한, 장치.

청구항 100

제 99 항에 있어서,

상기 제1 복원 계층에 대한 상기 제1 신경망 매개변수(13)를 상기 데이터 스트림 또는 별도의 데이터 스트림으로 인코딩하고,

상기 제1 및 제2 복원 계층에 대한 개별의 확률 컨텍스트를 사용하는 컨텍스트 적응 엔트로피 인코딩에 의해 상기 제2 복원 계층에 대한 상기 제2 신경망 매개변수(13)를 상기 데이터 스트림으로 인코딩하도록

구성되는, 장치.

청구항 101

제 99 항 또는 제 100 항에 있어서,

상기 신경망 매개변수(13)는 신경망 매개변수당 상기 제 1 복원 계층 신경망 매개변수 값 및 상기 제 2 복원 계층 신경망 매개변수 값의 매개변수 별 합 또는 매개변수 별 곱에 의해 복원 가능한, 장치.

청구항 102

제 99 항 내지 제 101 항 중 어느 한 항에 있어서, 상기 제 1 복원 계층 신경망 매개변수 값에 좌우되는 확률 모델을 사용하는 컨텍스트 적응 엔트로피 인코딩에 의해 상기 제 2 복원 계층 신경망 매개변수 값을 상기 데이터 스트림으로 인코딩하도록 구성되는, 장치.

청구항 103

제 99 항 내지 제 102 항 중 어느 한 항에 있어서,

컨텍스트 적응 엔트로피 인코딩하고,

상기 제 1 복원 계층 신경망 매개변수 값에 따라 확률 컨텍스트 세트의 집합 중에서 확률 컨텍스트 세트를 선택하고,

상기 제1 복원 계층 신경망 매개변수 값에 따라 상기 선택된 확률 컨텍스트 세트 중에서 사용될 확률 컨텍스트를 선택함으로써,

상기 제 2 복원 계층 신경망 매개변수 값을 상기 데이터 스트림으로 인코딩하도록 구성되는, 장치.

청구항 104

제 103 항에 있어서, 상기 확률 컨텍스트 세트의 집합은 3개의 확률 컨텍스트 세트를 포함하고, 상기 장치는,

상기 제 1 복원 계층 신경망 매개변수 값이 음수이면 상기 확률 컨텍스트 세트의 집합 중에서 제 1 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고,

상기 제 1 복원 계층 신경망 매개변수 값이 양수이면 상기 확률 컨텍스트 세트의 집합 중에서 제 2 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고,

상기 제 1 복원 계층 신경망 매개변수 값이 0이면 상기 확률 컨텍스트 세트의 집합 중에서 제 3 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하도록

구성되는, 장치.

청구항 105

제 103 항에 있어서, 상기 확률 컨텍스트 세트의 집합은 2개의 확률 컨텍스트 세트를 포함하고, 상기 장치는,

상기 제1 복원 계층 신경망 매개변수 값이 미리 결정된 값보다 큰 경우 상기 확률 컨텍스트 세트의 집합 중에서 제1 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고, 상기 제1 복원 계층 신경망 매개변수 값이 상기 미리 결정된 값보다 크지 않은 경우 상기 확률 컨텍스트 세트의 집합 중에서 제2 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하거나,

상기 제1 복원 계층 신경망 매개변수 값의 절대값이 상기 미리 결정된 값보다 큰 경우, 상기 확률 컨텍스트 세트의 집합 중에서 제1 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하고, 상기 제1 복원 계층 신경망 매개변수 값의 절대값이 상기 미리 결정된 값보다 크지 않은 경우, 상기 확률 컨텍스트 세트의 집합 중에서 제2 확률 컨텍스트 세트를 상기 선택된 확률 컨텍스트 세트로 선택하도록

구성되는, 장치.

청구항 106

데이터 스트림(14)으로부터 신경망(10)을 정의하는 신경망 매개변수(13)를 디코딩하는 방법(400)에 있어서, 상기 방법은,

현재 신경망 매개변수(13')에 대해, 이전 신경망 매개변수에 대해 상기 데이터 스트림(14)으로부터 디코딩된 양자화 인덱스(58)에 따라 복수(50)의 복원 레벨 세트(52) 중에서 복원 레벨 세트(48)를 선택하고(54)

상기 데이터 스트림(14)으로부터 상기 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩(420)하고, - 상기 양자화 인덱스(56)는 현재 신경망 매개변수에 대해 상기 선택된 복원 레벨 세트(48) 중 하나의 복원 레벨을 나타냄 - ,

상기 현재 신경망 매개변수(13')를 상기 현재 신경망 매개변수에 대한 상기 양자화 인덱스(56)에 의해 표시되는 상기 선택된 복원 레벨 세트(48)의 하나의 복원 레벨로 역양자화(62)함으로써

상기 신경망 매개변수(13)를 순차적으로 디코딩하는 단계

를 포함하는, 방법.

청구항 107

신경망을 정의하는 신경망 매개변수를 데이터 스트림으로 인코딩하는 방법(500)에 있어서, 상기 방법은,

현재 신경망 매개변수(13')에 대해, 이전에 인코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로 인코딩된 양자화 인덱스(58)에 따라 복수(50)의 복원 레벨 세트(52) 중에서 복원 레벨 세트(48)를 선택하고(54)

상기 현재 신경망 매개변수(13')를 상기 선택된 복원 레벨 세트(48)의 하나의 복원 레벨로 양자화(64)하고,

상기 현재 신경망 매개변수에 대한 양자화 인덱스(56)가 상기 데이터 스트림(14)으로 양자화되는 상기 하나의 복원 레벨을 나타내는 상기 현재 신경망 매개변수(13')에 대한 상기 양자화 인덱스(56)를 인코딩(530)함으로써,

상기 신경망 매개변수(13)를 순차적으로 인코딩하는

단계를 포함하는, 방법.

청구항 108

신경망(10)을 정의하는 신경망 매개변수(13)를 복원하기 위한 방법(600)에 있어서, 상기 방법은,

제1 복원 계층에 대한 제1 신경망 매개변수를 유도하여(610) 신경망 매개변수 당 제1 복원 계층 신경망 매개변수 값을 산출하는 단계(610),

신경망 매개변수(13) 당 제2 복원 계층 신경망 매개변수 값을 산출하기 위해 데이터 스트림으로부터 제2 복원 계층에 대한 제2 신경망 매개변수(13)를 디코딩하는 단계(620), 및

각 신경망 매개변수에 대해 상기 제1 복원 계층 신경망 매개변수 값과 상기 제2 복원 계층 신경망 매개변수 값을 결합함으로써 상기 신경망 매개변수(13)를 복원하는 단계(630)

를 포함하는, 방법.

청구항 109

신경망 매개변수(13) 당 제1 복원 계층 신경망 매개변수 값을 포함하는 제1 복원 계층에 대한 제1 신경망 매개변수(13)를 사용하여, 신경망(10)을 정의하는 신경망 매개변수(13)를 인코딩하는 방법(700)에 있어서, 상기 방법은,

제2 복원 계층에 대한 제2 신경망 매개변수(13)를, 신경망 매개변수(13)당 제 2 복원 계층 신경망 매개변수 값을 포함하는 데이터 스트림으로 인코딩(710)하는

단계를 포함하고,

상기 신경망 매개변수(13)는, 각각의 신경망 매개변수에 대해 상기 제1 복원 계층 신경망 매개변수 값과 상기 제2 복원 계층 신경망 매개변수 값을 결합함으로써 복원 가능한, 방법.

청구항 110

제 107 항 또는 제 109 항에 따른 방법에 의해 인코딩된 데이터 스트림.

청구항 111

프로그램이 하나 이상의 컴퓨터에서 실행될 때 제 106 항, 제 107 항, 제 108 항 또는 제 109 항에 따른 방법을 실행하기 위한 프로그램 코드를 갖는 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명에 따른 실시예는 신경망 매개변수에 대한 코딩 개념에 관한 것이다.

배경 기술

[0002] 가장 기본적인 형태의 신경망은 요소별 비선형 함수가 뒤따르는 아핀 변환 체인을 구성한다. 이들은 도 1과 같이 방향성 비순환 그래프로 나타낼 수 있다. 도 1은 신경망의 개략도, 여기에서는 예시적으로 2계층 피드 포워드 신경망을 도시한다. 즉, 도 1은 피드포워드 신경망의 그래프 표현을 도시한다. 특히, 이 2 계층 신경망은 4차원 입력 벡터를 실제 라인에 매핑하는 비선형 함수이다. 신경망은 4차원 입력 벡터에 따라 신경망의 입력인 입력 계층에 4개의 뉴런(10c)을 포함하고 은닉 계층에 5개의 뉴런(10c) 및 신경망의 출력을 형성하는 출력 층에 1개의 뉴런(10c)을 포함한다. 신경망은 서로 다른 또는 후속하는 계층의 뉴런을 연결하는, 뉴런 상호 연결부(11)를 더 포함한다. 뉴런 상호연결부(11)는 가중치와 연관될 수 있으며, 여기서 가중치는 서로 연결된 뉴런(10c) 간의 관계와 연관된다. 특히 가중치는 후속 계층으로 전달될 때 한 계층의 뉴런 활성화에 가중치를 부여하고, 여기에서 인바운드 가중치 활성화의 합은 선형 함수에 해당하는 후속 계층의 각 뉴런에서 다음에 비선형 함수에 해당하는 후속 계층의 각 뉴런/노드에서 가중치 합에 적용되는 비선형 스칼라 함수가 형성된다. 따라서 각 노드, 예를 들어 뉴런(10c)은 특정 값을 수반하며, 이 값은 예지, 예를 들어 뉴런 연결(11)의 각 가중치 값과 곱하여 다음 노드로 전달된다. 그러면 인입하는 모든 값이 단순히 집계된다.

[0003] 수학적으로, 도 1의 신경망은 다음과 같은 방식으로 출력을 계산한다:

[0004]
$$\text{output} = \sigma(W_2 \cdot \sigma(W_1 \cdot \text{input}))$$

[0005] 여기서 W_2 및 W_1 은 신경망 매개변수이다. 예를 들어 신경망 가중치 매개변수(예지 가중치) 및 시그마는 일부 비선형 함수이다. 예를 들어, 소위 컨볼루션 계층은 [1]에 설명된 것처럼 행렬-행렬 곱으로 캐스팅하여 사용할 수도 있다. 이제부터 주어진 입력에서 출력을 계산하는 절차를 추론이라고 한다. 또한 중간 결과를 은닉 계층 또는 은닉 활성화 값으로 부르며, 이는 예를 들어 위의 제 1 내적 + 비선형성의 계산과 같은, 선형 변환 + 요소별 비선형성을 구성한다.

[0006] 일반적으로 신경망에는 수백만 개의 매개변수가 있으므로, 표현하려면 수백 MB(예: 메가바이트)가 필요할 수 있다. 결과적으로, 그들의 추론 절차는 큰 행렬 간의 많은 내적 연산의 계산을 포함하기 때문에 실행하기 위해 높은 계산 자원이 필요하다. 따라서 이러한 내적을 수행하는 복잡성을 줄이는 것이 매우 중요하다.

[0007] 마찬가지로, 위에서 언급한 문제 외에도, 신경망의 많은 매개변수를 저장해야 하며 예를 들어 서버에서 클라이언트로 전송해야 할 수도 있다. 또한 연합 학습 환경, 또는 특정 수신자가 지불한 품질의 다른 단계에서 신경망 매개변수화를 제공하는 경우와 같이 점진적으로 신경망의 매개변수화에 대한 정보를 엔티티에 제공할 수 있거나, 추론을 위해 신경망을 사용할 때 처리할 수 있는 것이 유리하다.

발명의 내용

해결하려는 과제

[0008] 따라서, 예를 들어 압축과 같은 측면에서 보다 효율적인 신경망 매개변수의 효율적인 코딩을 위한 개념을 제공하는 것이 요망된다. 추가적으로 또는 대안적으로, 비트 스트림을 줄이고 이에 따라 신경망 매개변수에 대한 신호화 비용을 줄이는 것이 바람직하다.

[0009] 이 목적은 본 출원의 독립 청구항의 주제에 의해 달성된다.

[0010] 본 발명에 따른 추가 실시예는 본 출원의 종속항의 주제에 의해 정의된다.

과제의 해결 수단

- [0011] 본 발명의 제1 측면에 따른 실시예는, 이전 신경망 매개변수에 대한 데이터 스트림에서 디코딩된 양자화 인덱스에 따라 현재 신경망 매개변수에 대해 복수의 복원 레벨 세트 중에서 복원 레벨 세트를 선택함으로써 신경망 매개변수를 순차적으로 디코딩하도록 구성된, 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 디코딩하기 위한 장치를 포함한다. 또한, 장치는 데이터 스트림에서 현재 신경망 매개변수에 대한 양자화 인덱스를 디코딩함으로써 - 양자화 인덱스는 현재 신경망 매개변수에 대한 복원 레벨의 선택된 세트 중 하나의 복원 레벨을 나타냄 - 및 현재 신경망 매개변수에 대한 양자화 인덱스에 의해 표시되는 복원 레벨의 선택된 세트의 하나의 복원 레벨로 현재 신경망 매개변수를 역양자화함으로써, 신경망 매개변수를 순차적으로 디코딩하도록 구성된다.
- [0012] 본 발명의 제1 측면에 따른 추가 실시예는 이전에 인코딩된 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스에 따라 현재 신경망 매개변수에 대해 복수의 복원 레벨 세트 중에서 복원 레벨 세트를 선택함으로써, 신경망 매개변수를 순차적으로 인코딩하도록 구성된, 데이터 스트림으로 신경망을 정의하는 신경망 매개변수를 인코딩하기 위한 장치를 포함한다. 또한, 장치는 현재 신경망 매개변수를 선택된 복원 레벨 세트의 하나의 복원 레벨로 양자화함으로써, 및 현재 신경망 매개변수에 대한 양자화 인덱스가 데이터 스트림으로 양자화되는 하나의 복원 레벨을 나타내는 현재 신경망 매개변수에 대한 양자화 인덱스를 인코딩함으로써 신경망 매개변수를 순차적으로 인코딩하도록 구성된다.
- [0013] 본 발명의 제1 측면에 따른 추가 실시예는 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 디코딩하는 방법을 포함한다. 방법은 이전 신경망 매개변수에 대한 데이터 스트림에서 디코딩된 양자화 인덱스에 따라 현재 신경망 매개변수에 대해 복수의 복원 레벨 세트 중에서 복원 레벨 세트를 선택함으로써 신경망 매개변수를 순차적으로 디코딩하는 단계를 포함한다. 또한, 방법은 데이터 스트림에서 현재 신경망 매개변수에 대한 양자화 인덱스를 디코딩함으로써 - 상기 양자화 인덱스는 현재 신경망 매개변수에 대한 복원 레벨의 선택된 세트 중 하나의 복원 레벨을 나타냄 - 및 현재 신경망 매개변수에 대한 양자화 인덱스에 의해 표시되는 복원 레벨의 선택된 세트의 하나의 복원 레벨로 현재 신경망 매개변수를 역양자화함으로써 신경망 매개변수를 순차적으로 인코딩하는 단계를 포함한다.
- [0014] 본 발명의 제1 측면에 따른 추가 실시예는 신경망을 정의하는 신경망 매개변수를 데이터 스트림으로 인코딩하는 방법을 포함한다. 방법은 이전에 인코딩된 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스에 따라 현재 신경망 매개변수에 대해 복수의 복원 레벨 세트 중에서 복원 레벨 세트를 선택함으로써 신경망 매개변수를 순차적으로 인코딩하는 단계를 포함한다. 또한, 방법은 현재 신경망 매개변수를 선택된 복원 레벨 세트의 하나의 복원 레벨로 양자화함으로써, 및 현재 신경망 매개변수에 대한 양자화 인덱스가 데이터 스트림으로 양자화되는 하나의 복원 레벨을 나타내는 현재 신경망 매개변수에 대한 양자화 인덱스를 인코딩함으로써, 신경망 매개변수를 순차적으로 인코딩하는 단계를 포함한다.
- [0015] 본 발명의 제1 측면에 따른 실시예는 신경망 매개변수가 일정하지 않은 양자화기를 사용하지만, 신경망 매개변수를 코딩하는 동안에 이를 변경하여, 즉, 이전 또는 이전에 인코딩된 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩되거나 이에 인코딩된 양자화 인덱스에 따라 복원 레벨 세트를 선택함으로써, 보다 효율적으로 압축될 수 있다는 아이디어에 기초한다. 따라서 신경망 매개변수의 정렬된 세트를 참조할 수 있는 복원 벡터는 N 차원 신호 공간에서 더 조밀하게 패킹될 수 있으며, 여기서 N은 처리될 샘플 세트의 신경망 매개변수의 수를 나타낸다. 이러한 종속 양자화는 각각 복호화 장치의 복호화 및 역양자화 또는 인코딩 장치의 양자화 및 부호화에 사용될 수 있다.
- [0016] 본 발명의 제2 측면에 따른 실시예는 신경층에서 신경망의 계층화된 구성과 구별하기 위해 복원 계층이라고 불리는 단계별로 수행될 때, 이들 단계에서 제공되는 매개변수화가 신경망 매개변수별로 결합되어 다른 단계에 비해 개선된 신경망 매개변수화를 생성하는 경우 보다 효율적인 신경망 코딩을 달성할 수 있다는 아이디어에 기초한다. 따라서, 신경망을 정의하는 신경망 매개변수 복원 장치는, 제1 복원 계층이 신경망 매개변수당 제1 복원 계층 신경망 매개변수 값을 산출하기 위해, 제 1 신경망 매개변수, 예를 들어, 제 1 복원 계층 신경망 매개변수를 유도한다.
- [0017] 제 1 신경망 매개변수는 예를 들어 연합 학습 프로세스 중에 이전에 전송될 수 있다. 더욱이, 제1 신경망 매개변수는 제1 복원 계층 신경망 매개변수 값일 수 있다. 또한, 장치는 예를 들어, 제2 신경망 매개변수, 제 2 복원 계층 신경망 매개변수를 디코딩하도록 구성되어, 이들을 예를 들어 최종 신경망 매개변수와 구별하므로, 데

이터 스트림으로부터의 제2 복원 계층이 신경망 매개변수당 제2 복원 계층 신경망 매개변수 값을 산출할 수 있다. 제 2 신경망 매개변수는 신경망 표현 측면에서 자체 포함된 의미가 없을 수 있지만, 제 1 표현 계층의 매개변수와 결합될 때, 신경망 표현, 즉 예를 들어 최종 신경망 매개변수로 이어질 수 있다. 또한, 장치는 각각의 신경망 매개변수에 대해 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써 신경망 매개변수를 복원하도록 구성된다.

[0018] 본 발명의 제2 측면에 따른 추가 실시예는 신경망 매개변수당 제1 복원 계층 신경망 매개변수 값을 포함하는 제1 복원 계층에 대한 제1 신경망 매개변수를 사용함으로써 신경망을 정의하는 신경망 매개변수를 인코딩하기 위한 장치를 포함한다. 또한, 장치는 제2 복원 계층에 대한 제2 신경망 매개변수를 신경망 매개변수당 제 2 복원 계층 신경망 매개변수 값을 포함하는 데이터 스트림으로 인코딩하도록 구성되며, 여기서 신경망 매개변수는 각각의 신경망 매개변수에 대해 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써 복원 가능하다.

[0019] 본 발명의 제2 측면에 따른 추가 실시예는 신경망을 정의하는 신경망 매개변수를 복원하기 위한 방법을 포함한다. 방법은 예를 들어 연합 학습 프로세스 동안 이전에 전송될 수 있으며, 예를 들어, 제1 복원 계층이 신경망 매개변수당 제1 복원 계층 신경망 매개변수 값을 산출하도록 제1 복원 계층 신경망 매개변수라고 부를 수 있는, 제1 신경망 매개변수를 유도하는 단계를 포함한다.

[0020] 또한, 방법은 예를 들어, 최종 복원된 신경망 매개변수와 구별하기 위해서 예를 들어 제 2 복원 계층 신경망 매개변수라고 할 수 있는 제2 신경망 매개변수를 디코딩하여, 데이터 스트림으로부터 제2 복원 계층이 신경망 매개변수당 제2 복원 계층 신경망 매개변수 값을 산출하도록 하고, 방법은 각각의 신경망 매개변수에 대해 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써 신경망 매개변수를 복원하는 단계를 포함한다. 제 2 신경망 매개변수는 신경 표현 측면에서 자체 포함된 의미가 없을 수 있지만, 제 1 표현 계층의 매개변수와 결합될 때, 신경망 표현, 즉 예를 들어 최종 신경망 매개변수로 이어질 수 있다.

[0021] 본 발명의 제2 측면에 따른 추가 실시예는 신경망 매개변수당 제1 복원 계층 신경망 매개변수 값을 포함하는 제1 복원 계층에 대한 제1 신경망 매개변수를 사용함으로써 신경망을 정의하는 신경망 매개변수를 인코딩하는 방법을 포함한다. 방법은 제2 복원 계층에 대한 제2 신경망 매개변수를 신경망 매개변수당 제 2 복원 계층 신경망 매개변수 값을 포함하는 데이터 스트림으로 인코딩하는 단계를 포함하고, 여기서 신경망 매개변수는 각각의 신경망 매개변수에 대해 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써, 복원 가능하다.

[0022] 본 발명의 제2 측면에 따른 실시예는 예를 들어 신경망 매개변수에 의해 정의된 신경망은 예를 들어 복원 계층, 예를 들어 기본 계층 및 향상 계층과 같은 하위 계층을 사용하여 비트스트림에서 적은 양의 데이터를 사용하여, 효율적으로 압축 및/또는 전송될 수 있다는 아이디어에 기초한다. 복원 계층은 각각의 신경망 매개변수에 대해 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써 신경망 매개변수가 복원 가능하도록 정의될 수 있다. 이 분포는 신경망 매개변수의 효율적인 코딩, 예를 들어, 인코딩 및/또는 디코딩, 및/또는 전송을 가능하게 한다. 따라서, 제2 복원 계층에 대한 제2 신경망 매개변수는 데이터 스트림으로 별도로 인코딩 및/또는 전송될 수 있다.

도면의 간단한 설명

[0023] 도면은 반드시 축척에 맞춰진 것은 아니며, 일반적으로 본 발명의 원리를 설명하는 데 중점을 둔다. 다음 설명에서, 본 발명의 다양한 실시예는 다음 도면을 참조하여 설명된다:

도 1은 본 발명의 실시예와 함께 사용될 수 있는 2 계층 피드-포워드 신경망의 예시의 개략도를 도시한다.

도 2는 일 실시예에 따른, 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 디코딩하기 위한 장치 내에서 수행되는 역양자화를 위한 개념의 개략도를 도시한다.

도 3은 일 실시예에 따른, 신경망 매개변수를 데이터 스트림으로 인코딩하기 위한 장치 내에서 수행되는 양자화를 위한 개념의 개략도를 도시한다.

도 4는 일 실시예에 따른, 신경망을 정의하는 신경망 매개변수를 복원하기 위한 장치 내에서 수행되는 디코딩을 위한 개념의 개략도를 도시한다.

도 5는 일 실시예에 따른, 신경망을 정의하는 신경망 매개변수를 복원하기 위한 장치 내에서 수행되는 인코딩

개념의 개략도를 도시한다.

도 6은 본 발명에 따른 실시예와 함께 사용하기 위한 신경망 매개변수에 대한 복원 계층을 사용하는 개념의 개략도를 도시한다.

도 7은 본 발명의 실시예에 따른 균일 복원 양자화기의 예시의 개략도를 도시한다.

도 8은 본 발명의 실시예에 따른 2개의 가중치 매개변수의 단순한 경우에 대해 허용 가능한 복원 벡터의 위치의 예를 도시한다.

도 9는 본 발명의 실시예에 따라 단일 양자화 단계 크기 Δ 에 의해 완전히 결정되는, 두 세트의 복원 레벨을 사용한 종속 양자화의 예를 도시한다.

도 10은 본 발명의 실시예에 따른, 신경망 매개변수에 대한 복원 프로세스의 바람직한 예를 나타내는 의사 코드의 예를 도시한다.

도 11은 본 발명의 실시예에 따른, 복원 레벨 세트를 2개의 하위 세트로 분할하는 예를 도시한다.

도 12는 실시예에 따른 계층에 대한 신경망 매개변수의 복원 프로세스에 대한 바람직한 예를 나타내는 의사 코드의 예를 도시한다.

도 13은 본 발명의 실시예에 따른, 상태와 연관된 양자화 세트를 지정하는 상태 전이 표 sttab 및 표 setId에 대한 바람직한 예를 도시한다.

도 14는 본 발명의 실시예에 따른, 상태와 연관된 양자화 세트를 지정하는, 상태 전이 표 sttab 및 표 setId에 대한 바람직한 예를 도시한다.

도 15는 본 발명의 실시예에 따른, 0과 동일한 양자화 인덱스가 상태 전이 및 종속 스칼라 양자화에서 제외되는, 신경망 매개변수 레벨에 대한 대안적인 복원 프로세스를 예시하는 의사 코드를 도시한다;

도 16은 본 발명의 실시예에 따른, 격자 구조로서의 종속 스칼라 양자화에서의 상태 전이의 예를 도시한다.

도 17은 본 발명의 실시예에 따른 기본 격자 셀의 예를 도시한다.

도 18은 본 발명의 실시예에 따른, 8개의 신경망 매개변수의 종속 스칼라 양자화를 위한 격자의 예를 도시한다.

도 19는 본 발명의 실시예에 따른, (라그랑주 비용 측정 $D + \lambda \cdot R$ 과 같은) 비용 측정을 최소화하는 양자화 인덱스의 시퀀스 (또는 블록)를 결정하기 위해 이용될 수 있는 예시적인 격자 구조를 도시한다.

도 20은 본 발명의 실시예에 따른, 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 디코딩하는 방법의 블록도를 도시한다.

도 21은 본 발명의 실시예에 따른, 신경망을 정의하는 신경망 매개변수를 데이터 스트림으로 인코딩하는 방법의 블록도를 도시한다.

도 22는 본 발명의 실시예에 따른, 신경망을 정의하는 신경망 매개변수를 복원하기 위한 방법의 블록도를 도시한다; 및

도 23은 본 발명의 실시예에 따른, 신경망을 정의하는 신경망 매개변수를 인코딩하기 위한 방법의 블록도를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0024] 동일하거나 동등한 요소 또는 동일하거나 동등한 기능을 갖는 요소는 상이한 도면에서 발생하더라도 동일하거나 동등한 참조 번호로 이하의 설명에서 표시된다.

[0025] 다음 설명에서, 본 발명의 실시예에 대한 보다 철저한 설명을 제공하기 위해 복수의 세부사항이 제시된다. 그러나, 본 발명의 실시예가 이러한 특정 세부사항 없이 실시될 수 있음은 당업자에게 명백할 것이다. 다른 예에서, 잘 알려진 구조 및 장치는 본 발명의 실시예를 모호하게 하는 것을 피하기 위해 상세하기보다는 블록도 형태로 도시된다. 또한, 이후에 설명되는 상이한 실시예들의 특징들은 특별히 달리 언급되지 않는 한 서로 결합될 수 있다.

[0026] 설명은 본 출원의 일부 실시예의 제시로 시작한다. 이 설명은 매우 일반적이지만 독자에게 본 출원의 실시예가

기반으로 하는 기능의 개요를 제공한다. 이어서, 이러한 기능에 대한 보다 상세한 설명이 실시예에 대한 동기 및 위에서 설명한 효율성 이득을 달성하는 방법과 함께 제공된다. 세부 사항은 지금 설명된 실시예와 개별적으로 및 조합하여 조합할 수 있다.

[0027] 도 2는 일 실시예에 따른 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 디코딩하기 위한 장치 내에서 수행되는 역양자화를 위한 개념의 개략도를 도시한다. 신경망은 예를 들어 상호 연결된 층의 뉴런 사이의 뉴런 상호 연결을 갖는, 복수의 상호 연결된 신경망 층을 포함한다. 도 2는 데이터 스트림(14)에서 예를 들어 인코딩된 신경망 매개변수(13)에 대한 양자화 인덱스(56)를 도시한다. 따라서, 신경망 매개변수(13)는 신경망 사이의 가중치와 관련하여 신경망을 정의하거나 매개변수화할 수 있다.

[0028] 장치는 신경망 매개변수(13)를 순차적으로 디코딩하도록 구성된다. 이 순차적 처리 동안, 양자화기(복원 레벨 세트)가 변경된다. 이 변형을 통해 더 적은(또는 더 적은 밀도) 레벨의 양자화기를 사용하는 것을 가능하게 하므로, 더 작은 양자화 인덱스가 코딩될 수 있게 하고, 이 때 필요한 코딩 비트레이트와 비교하여 이러한 양자화로 인한 신경망 표현의 품질은 일정한 양자화기를 사용하는 것에 비해 개선된다. 세부 사항은 나중에 설정된다. 특히, 장치는 이전 신경망 매개변수에 대해 데이터 스트림(14)으로부터 디코딩된 양자화 인덱스(58)에 따라, 현재 신경망 매개변수(13')에 대해, 복수(50)의 복원 레벨 세트(52)(세트 0, 세트 1) 중 복원 레벨 세트(48)(선택된 세트)를 선택(54)(복원 레벨 선택)함으로써, 신경망 매개변수(13)를 순차적으로 디코딩한다.

[0029] 게다가, 장치는 데이터 스트림(14)으로부터 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩함으로써, - 양자화 인덱스(56)는 현재 신경망 매개변수에 대한 복원 레벨의 선택된 세트(48) 중 하나의 복원 레벨을 나타냄 -, 및 현재 신경망 매개변수(13')를 현재 신경망 매개변수에 대한 양자화 인덱스(56)에 의해 표시되는 복원 레벨의 선택된 세트(48) 중 하나의 복원 레벨로 역양자화(62)함으로써, 신경망 매개변수(13)를 순차적으로 디코딩하도록 구성된다.

[0030] 디코딩된 신경망 매개변수(13)는 일 예로서 행렬(15a)로 표현된다. 행렬은 신경망의 뉴런 상호연결의 가중치와 관련될 수 있는 역직렬화된(20b) 신경망 매개변수(13)를 포함할 수 있다.

[0031] 선택적으로, 복원 레벨 세트(52)의 복수(50) 중, 본 명세서에서 때때로 양자화기라고도 하는 복원 레벨 세트(52)의 수는 예를 들어 도 2에 도시된 바와 같이, 2개, 즉 세트 0 및 세트 1일 수 있다.

[0032] 더욱이, 장치는 예를 들어 Δ 또는 Δk 로 표시되는, 미리 결정된 양자화 단계 크기(QP)에 의해, 복수(50)의 복원 레벨 세트(52)(예를 들어, 세트 0, 세트 1)를 매개변수화(60)하고, 데이터 스트림(14)으로부터 미리 결정된 양자화 단계 크기에 대한 정보를 유도하도록 구성된다. 따라서, 실시예에 따른 디코더는 가변 스텝 사이즈(QP)에 적응할 수 있다.

[0033] 또한, 실시예들에 따르면, 신경망은 하나 이상의 NN 계층을 포함할 수 있고 장치는 각각의 NN 계층에 대해 데이터 스트림(14)으로부터 각각의 NN 계층에 대한 미리 결정된 양자화 단계 크기(QP)에 대한 정보를 유도하고, 각각의 NN 계층에 대해, 개별 NN 계층에 속하는 신경망 매개변수를 역양자화하는 데 사용되도록 각각의 NN 계층에 대해 유도된 미리 결정된 양자화 단계 크기를 사용하여 복수의 복원 레벨 세트(52)(50)를 매개변수화하도록 구성될 수 있다. NN 계층에 대한 스텝 크기 및 그에 따른 복원 레벨 세트(52)의 적응은 코딩 효율을 개선할 수 있다.

[0034] 추가 구현예에 따르면, 장치는 현재 신경망 매개변수(13')에 대해, 이전에 디코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로부터 디코딩된 양자화 인덱스(58)의 이진화의 LSB(예를 들어, 최하위 비트) 부분 또는 이전에 디코딩된 빈(예를 들어, 이진 결정)에 따라, 복수(50)의 복원 레벨 세트(52) 중 복원 레벨의 세트(48)를 선택(54)하도록 구성될 수 있다. LSB 비교는 낮은 계산 비용으로 수행될 수 있다. 특히, 상태 전이가 사용될 수 있다.

[0035] 선택(54)은 현재 신경망 매개변수(13')에 대해 현재 신경망 매개변수(13')와 연관된 상태에 따라 복수의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 결정함으로써, 및 직전 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상태를 업데이트함으로써, 상태 전이 프로세스에 의해 현재 신경망 매개변수(13')에 대해 복수(50)의 복원 레벨 세트(52) 중에서 양자화 레벨 세트(48)를 선택하도록 수행될 수 있다. 예를 들어, 전이 표의 사용에 의한 상태 전이 이외의 대안적 접근법이 또한 사용될 수 있고 아래에 설명되어 있다.

[0036] 추가적으로 또는 대안적으로, 장치는 예를 들어 현재 신경망 매개변수(13')에 대해 이전에 디코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로부터 디코딩된 양자화 인덱스(58)의 이진 함수의 결과에 따라, 복수(50)의

복원 레벨 세트(52) 중 복원 레벨 세트(48)를 선택(54)하도록 구성될 수 있다. 이진 함수는 예를 들어, 비트 단위 "and" 연산을 사용하는, 예를 들어 패리티 검사로, 양자화 인덱스(58)가 짝수 또는 홀수를 나타내는지 여부를 시그널링한다. 이것은 양자화 인덱스(58)를 인코딩하는 데 사용되는 복원 레벨의 세트(48)에 대한 정보를 제공할 수 있으므로, 복원 레벨의 미리 결정된 순서 때문에, 현재 신경망 매개변수(13')을 인코딩하는 데 사용되는 복원 레벨 세트에 대해, 대응하는 인코더에서 사용되는 세트를 결정한다. 패리티는 앞에서 언급한 상태 전이에 사용될 수 있다.

[0037] 또한, 실시예들에 따르면, 장치는 예를 들어 현재 신경망 매개변수(13')에 대해 이전에 디코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로부터 디코딩된 양자화 인덱스(58)의 패리티에 따라 복원 레벨 세트(52)의 복수(50) 중 복원 레벨 세트(48)를 선택(54)하도록 구성될 수 있다. 패리티 검사는 낮은 계산 비용으로 수행될 수 있다. 비트 단위 "and" 연산을 사용한다.

[0038] 선택적으로, 장치는 신경망 매개변수(13)에 대해 양자화 인덱스(56)를 디코딩하고 신경망 매개변수(13) 간의 공통 순차 순서(14')를 따라 신경망 매개변수(13)의 역양자화를 수행하도록 구성될 수 있다. 즉, 두 작업에 동일한 순서를 사용할 수 있다.

[0039] 도 3은 일 실시예에 따라 신경망 매개변수를 데이터 스트림으로 인코딩하기 위한 장치 내에서 수행되는 양자화를 위한 개념의 개략도를 도시한다. 도 3은 신경망 층(10a, 10b)을 포함하는 신경망(NN)(10)을 도시하고, 여기서 층들은 뉴런(10c)을 포함하고 상호연결된 층의 뉴런은 뉴런 상호연결부(11)를 통해 상호연결된다. 예를 들어, NN 계층(p-1)(10a)와 NN 계층(p)(10b)가 도시되고, 여기서 p는 NN 계층에 대한 인덱스이며, $1 \leq p \leq \text{NN의 계층 수}$ 이다. 신경망은 선택적으로 신경망(10)의 뉴런 상호접속부(11)의 가중치와 관련될 수 있는 신경망 매개변수(13)에 의해 정의되거나 매개변수화된다. 도 1의 은닉 계층의 뉴런(10c)은 도 3의 계층 p(A, B, C, ...)의 뉴런을 나타낼 수 있고, 도 1의 입력 계층의 뉴런은 도 3에 도시된 계층 p-1(a, b, c, ...)의 뉴런을 나타낼 수 있다. 신경망 매개변수(13)는 도 1의 뉴런 상호연결부(11)의 가중치와 관련될 수 있다.

[0040] 상이한 층의 뉴런(10c)의 관계는 신경망 매개변수(13)의 행렬(15a)에 의해 도 1에 표시된다. 예를 들어, 네트워크 매개변수(13)가 뉴런 상호연결(11)의 가중치와 관련된 경우, 행렬(15a)는 예를 들어 행렬 요소가 상이한 층의 뉴런(10c) 사이의 가중치를 나타내도록 구조화될 수 있다(예: 계층 p-1의 경우 a, b, ... 및 계층 p의 경우 A, B, ...).

[0041] 장치는 예를 들어 직렬(20a)(직렬화)에서 신경망 매개변수(13)를 순차적으로 인코딩하도록 구성된다. 이 순차 처리 동안, 양자화(복원 레벨 세트)가 변경된다. 이 변형을 통해 더 적은(또는 더 적은 밀도) 레벨의 양자화기를 사용할 수 있으므로, 더 작은 양자화 인덱스를 코딩할 수 있으며, 여기서 필요한 코딩 비트레이트와 비교하여 이러한 양자화로 인한 신경망 표현의 품질은 일정한 양자화기를 사용하는 것과 비교하여 개선된다. 세부 사항은 나중에 설정된다. 특히, 장치는 이전에 인코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로 인코딩된 양자화 인덱스(58)에 따라, 현재 신경망 매개변수(13')에 대해, 복수(50)의 복원 레벨 세트(52)의 중 복원 레벨 세트(48)를 선택(54)함으로써, 신경망 매개변수(13)을 순차적으로 인코딩한다.

[0042] 게다가, 장치는 현재 신경망 매개변수(13')를 복원 레벨의 선택된 세트(48)중 하나의 복원 레벨로 양자화(64(Q))함으로써, 및 현재 신경망 매개변수에 대한 양자화 인덱스(56)가 데이터 스트림(14)으로 양자화되는 하나의 복원 레벨을 나타내는 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 인코딩함으로써, 신경망 매개변수(13)를 순차적으로 인코딩하도록 구성된다. 선택적으로, 복수(50)의 복원 레벨 세트(52) 중, 본 명세서에서 때때로 양자화기라고도 하는 복원 레벨 세트(52)의 수는 2로서, 예를 들어, 세트 0과 세트 1을 사용하여 표시한 바와 같다.

[0043] 실시예들에 따르면, 도 3에 도시된 바와 같이, 장치는, 예를 들어, 미리 결정된 양자화 단계 크기(QP)를 통해 복수(50)의 복원 레벨 세트(52)를 매개변수화(60)하고 미리 결정된 양자화 단계 크기에 대한 정보를 데이터 스트림(14)에 삽입하도록 구성될 수 있다. 이것은 예를 들어 양자화 효율을 개선하기 위해 적응적 양자화를 가능하게 할 수 있고, 여기서 신경망 매개변수(13)가 인코딩되는 방식의 변경은 미리 결정된 양자화 단계 크기에 대한 정보와 함께 디코더에 전달될 수 있다. By는 미리 정해진 양자화 단계 크기(QP)를 사용하여 정보 전송을 위한 데이터의 양을 줄일 수 있다.

[0044] 또한, 실시예들에 따르면, 신경망(10)은 하나 이상의 NN 층(10a, 10b)을 포함할 수 있고 장치는 각 NN 계층(p; p-1)에 대해 각 NN 계층에 해한 미리 결정된 양자화 단계 크기(QP)에 대한 정보를 데이터 스트림(14)에 삽입하고, 각각의 NN 계층에 대해, 각 NN 계층에 속하는 신경망 매개변수를 양자화하는 데 사용되도록 각각의 NN 계층

에 대해 유도된 미리 결정된 양자화 단계 크기를 사용하여 복수(50)의 복원 레벨 세트(52)를 매개변수화하도록 구성될 수 있다. 상술한 바와 같이, 예를 들어 NN 계층 또는 NN 계층의 특성에 따른, 양자화의 적응은 양자화 효율을 향상시킬 수 있다.

[0045] 선택적으로, 장치는 이전에 인코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로 인코딩된 양자화 인덱스(58)의 이진화의 LSB 부분 또는 이전에 인코딩된 빈에 따라 현재 신경망 매개변수(13')에 대해, 복수(50)의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 선택(54)하도록 구성될 수 있다. LSB 비교는 낮은 계산 비용으로 수행될 수 있다.

[0046] 유사하게, 도 2에 설명된 디코딩 장치에 상태 전이가 사용될 수 있다. 선택(54)은 현재 신경망 매개변수(13')와 관련된 상태에 따라, 현재 신경망 매개변수(13')에 대해 복수(50)의 복원 레벨 세트(52) 중에서 복원 레벨 세트(48)를 결정하고 직전의 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상태를 업데이트함으로써, 상태 전환 프로세스를 통해 현재 신경망 매개변수(13')에 대해 복수(50)의 복원 레벨 세트(52) 중 양자화 레벨의 세트(48)의 선택이 수행될 수 있다. 예를 들어, 전이 표의 사용에 의한 상태 전이 이외의 대안적 접근법이 또한 사용될 수 있고 아래에 설명되어 있다.

[0047] 추가적으로 또는 대안적으로, 장치는 이전에 인코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로 인코딩된 양자화 인덱스(58)의 이진 함수의 결과에 따라, 현재 신경망 매개변수(13')에 대해, 복원 레벨 세트(52)의 복수(50) 중 복원 레벨 세트(48)를 선택(54)하도록 구성될 수 있다. 이진 함수는 비트 단위 "and" 연산을 사용하는 예를 들어 패리티 검사일 수 있으며, 양자화 인덱스(58)가 짝수 또는 홀수를 나타내는지 여부를 시그널링한다. 이것은 양자화 인덱스(58)를 인코딩하는 데 사용되는 복원 레벨의 세트(48)에 대한 정보를 제공할 수 있으므로, 예를 들어 복원 레벨의 미리 결정된 순서 때문에, 현재 신경망 매개변수(13')에 대한 복원 레벨 세트(48)를 결정하므로, 예를 들어 대응하는 디코더가 미리 결정된 순서 때문에 복원 레벨의 대응하는 세트(48)를 선택할 수 있도록 한다. 패리티는 앞에서 언급한 상태 전이에 사용될 수 있다.

[0048] 또한, 실시예들에 따르면, 장치는, 이전에 인코딩된 신경망 매개변수에 대해 데이터 스트림(14)으로 인코딩된 양자화 인덱스(56)의 패리티에 따라, 예를 들어 현재 신경망 매개변수(13')에 대해, 복원 레벨 세트(52)의 복수(50) 중 양자화 레벨 세트(48)를 선택(54)하도록 구성될 수 있다. 예를 들어 비트 단위 "and" 연산을 사용하여, 낮은 계산 비용으로 패리티 검사를 수행할 수 있다.

[0049] 선택적으로, 장치는 신경망 매개변수(13)에 대한 양자화 인덱스(56)를 인코딩하고 신경망 매개변수(13) 사이의 공통 순차 순서(14')를 따라 신경망 매개변수(13)의 양자화를 수행하도록 구성될 수 있다. 즉, 두 작업에 동일한 순서를 사용할 수 있다.

[0050] 도 4는 일 실시예에 따른 양자화된 신경망 매개변수를 산술 디코딩하기 위한 개념의 개략도를 도시한다. 이는 도 2의 장치 내에서 사용될 수 있다. 따라서 도 4는 도 2의 가능한 확장으로 볼 수 있다. 이것은 예를 들어 이진 산술 코딩을 사용하여 선택적 예제로 표시된 바와 같이, 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)가 산술 코딩을 사용하여 도 4의 장치에 의해 디코딩되는 데이터스트림(14)을 도시한다. 예를 들어, 화살표(123)로 표시된 바와 같이, 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨 세트(48)에 의존하는, 특정 컨텍스트에 의해 정의된 확률 모델이 사용된다. 이하 상세하게 설정한다.

[0051] 도 2와 관련하여 설명된 바와 같이, 현재 신경망 매개변수(13')에 대해 현재 신경망 매개변수(13')와 연관된 상태에 따라 복수의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 결정함으로써, 및 직전 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상태를 업데이트함으로써, 상태 전이 프로세스에 의해 복수(50)의 복원 레벨 세트(52) 중에서 양자화 레벨 세트(48)를 선택하는 선택(54)이 현재 신경망 매개변수(13')에 대해 수행된다. 따라서 상태는 현재 신경망 매개변수(13')를 인코딩/디코딩하는 데 사용되는 복원 레벨 세트(48)에 대한 유사 포인터로서, 이것은 복원 세트의 수에 해당하는 숫자 상태만 구별하는 것으로 더 세밀하게 업데이트되므로, 유사 상태는 과거 신경망 매개변수 또는 과거 양자화 인덱스의 메모리 역할을 한다. 따라서 상태는 신경망 매개변수(13)를 인코딩/디코딩하는 데 사용되는 복원 레벨 세트의 순서를 정의한다. 도 4에 따르면, 예를 들어, 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)는 현재 신경망 매개변수의 상태(13')에 의존(122)하는 확률 모델을 사용하는 산술 코딩을 사용하여 데이터 스트림(14)에서 디코딩된다. 상태에 따라 확률 모델을 적용하면 확률 모델 추정이 더 나을 수 있으므로 코딩 효율이 향상될 수 있다. 또한, 상태에 기초한 적응은 적은 추가 데이터의 양이 전송되면 계산상 효율적인 적응을 가능하게 할 수 있다.

- [0052] 추가 구현예에 따르면, 장치는 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈(84)에 대해 현재 신경망 매개변수(13')에 대한 상태에 의존하는(122) 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 데이터 스트림(14)으로부터 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩하도록 구성될 수 있다.
- [0053] 추가적으로 또는 대안적으로, 장치는 확률 모델의 의존성이 의존성을 사용하여 신경망 매개변수에 대한 컨텍스트 세트 중에서 하나의 컨텍스트(87)의 선택(103)(유도)을 포함하도록 구성될 수 있고, 각 컨텍스트는 관련된 미리 결정된 확률 모델을 갖는다. 더 나은 확률 추정을 사용할수록, 압축은 더 효율적이다. 확률 모델은 예를 들어, 컨텍스트 적응형(이진) 산술 코딩을 사용하여 업데이트될 수 있다.
- [0054] 선택적으로, 장치는 각각의 컨텍스트를 사용하여 산술적으로 코딩된 양자화 인덱스에 기초하여 컨텍스트 각각과 연관된 미리 결정된 확률 모델을 업데이트하도록 구성될 수 있다. 따라서 컨텍스트의 확률 모델은 실제 통계에 맞게 조정된다.
- [0055] 또한, 장치는 양자화 인덱스의 이진화의 적어도 하나의 빈에 대해 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨의 세트(48)에 의존하는 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 데이터 스트림(14)으로부터 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩하도록 구성될 수 있다.
- [0056] 선택적으로, 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)가 0과 같거나 같지 않은 것을 나타내는 유의성 빈을 포함할 수 있다. 추가적으로 또는 대안적으로, 적어도 하나의 빈은 0보다 크거나 0보다 작은 현재 신경망 매개변수의 양자화 인덱스(56)를 나타내는 부호 빈을 포함할 수 있다. 뿐만 아니라, 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)의 절대값이 X보다 크거나 작음을 나타내는 X보다 큰 빈을 포함할 수 있고, 여기서 X는 0보다 큰 정수이다.
- [0057] 이하, 도 5는 도 4에서 설명한 디코딩 개념의 대응물을 설명할 수 있다. 따라서 모든 설명 및 이점은 인코딩에 대한 다음 개념의 측면에 적절하게 적용될 수 있다.
- [0058] 도 5는 일 실시예에 따른 신경망 매개변수를 산술적으로 인코딩하기 위한 개념의 개략도를 도시한다. 이것은 도 3의 장치 내에서 사용될 수 있다. 따라서 도 5는 도 3의 가능한 확장으로 볼 수 있다. 이것은 예를 들어 이진 산술 코딩을 사용하는 것과 같이 선택적 예제로 도시되는 바와 같이, 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)가 산술 코딩을 사용하여 도 3의 장치에 의해 인코딩되는 데이터 스트림(14)을 도시한다. 화살표(123)로 표시된 바와 같이 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨 세트(48)에 좌우하는, 예를 들어, 특정 컨텍스트에 의해 정의되는 확률 모델이 사용된다. 이하 상세하게 설정된다.
- [0059] 도 3과 관련하여 설명된 바와 같이, 현재 신경망 매개변수(13')에 대해 현재 신경망 매개변수(13')와 연관된 상태에 따라 복수(50)의 복원 레벨 세트(52) 중 양자화 레벨 세트(48)를 결정함으로써, 및 직전 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상태를 업데이트함으로써, 상태 전이 프로세스에 의해 복수(50)의 복원 레벨 세트(52) 중에서 양자화 레벨 세트(48)를 선택하는 선택(54)이 현재 신경망 매개변수(13')에 대해 수행된다.
- [0060] 따라서 상태는 현재 신경망 매개변수(13')를 인코딩/디코딩하는 데 사용되는 복원 레벨 세트(48)에 대한 유사 포인터로, 이것은 복원 세트의 수에 해당하는 숫자 상태만 구별하므로 더 세밀하게 업데이트되므로, 유사 상태가 과거 신경망 매개변수 또는 과거 양자화 인덱스의 메모리 역할을 하도록 한다. 따라서 상태는 신경망 매개변수(13)를 인코딩/디코딩하는 데 사용되는 복원 레벨 세트의 순서를 정의한다.
- [0061] 또한, 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)는 현재 신경망 매개변수(13')에 대한 상태에 의존하는 확률 모델을 사용하여 산술 코딩을 사용하여 데이터 스트림(14)으로 인코딩될 수 있다.
- [0062] 도 3에 따르면, 예를 들어 양자화 인덱스(56)는 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈(84)에 대해 현재 신경망 매개변수(13')에 대한 상태에 의존하는 확률 모델을 사용함으로써 이진 산술 코딩을 이용하여 현재 신경망 매개변수(13')에 대해 데이터 스트림(14)으로 인코딩된다. 상태에 따라 확률 모델을 적용하면 확률 모델이 확률 모델 추정이 더 나을 수 있으므로 코딩 효율이 향상될 수 있다. 또한, 상태에 기반한 적응은 전송된 추가 데이터의 양이 적은 계산상 효율적인 적응을 가능하게 할 수 있다.
- [0063] 추가적으로 또는 대안적으로, 장치는 확률 모델의 의존성이 의존성을 사용하여 신경망 매개변수에 대한 컨텍스트 세트 중에서 하나의 컨텍스트(87)의 선택(103)(유도)을 포함하도록 구성될 수 있고, 각 컨텍스트는 관련된 미리 결정된 확률 모델을 갖는다.
- [0064] 선택적으로, 장치는 각각의 컨텍스트를 사용하여 산술적으로 코딩된 양자화 인덱스에 기초하여 컨텍스트 각각과

연관된 미리 결정된 확률 모델을 업데이트하도록 구성될 수 있다.

- [0065] 더욱이, 장치는 양자화 인덱스의 이진화의 적어도 하나의 빈에 대해 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨의 세트(48)에 의존하는 확률 모델을 사용함으로써 이진 산술 코딩을 사용하여 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 데이터 스트림(14)으로 인코딩하도록 구성될 수 있다. 이진 산술 코딩을 사용하기 위해 양자화 인덱스(56)는 이진화될 수 있다.
- [0066] 선택적으로, 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)가 0과 같거나 같지 않은 것을 나타내는 유의성 빈을 포함할 수 있다. 추가적으로 또는 대안적으로, 적어도 하나의 빈은 0보다 크거나 0보다 작은 현재 신경망 매개변수의 양자화 인덱스(56)를 나타내는 부호 빈을 포함할 수 있다. 또한, 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)의 절대값이 X보다 크거나 작음을 나타내는 X보다 큰 빈을 포함할 수 있고, 여기서 X는 0보다 큰 정수이다.
- [0067] 다음에 설명되는 실시예는 신경망의 매개변수화가 단계 또는 복원 계층에서 코딩되는 본 출원의 다른 측면에 집중하므로, NN 매개변수당, 각 단계의 하나의 값은, 적어도 하나는 그 자체가 신경망의 합리적인 표현을 나타낼 수 있는, 기여 단계들 중 하나로 향상되지만, 품질은 낮은 신경망의 개선/향상된 표현을 산출하도록 결합되는 것이 요구되며, 후자의 가능성은 현재 측면에서 필수 사항은 아니다.
- [0068] 도 6은 본 발명에 따른 실시예와 함께 사용하기 위한 신경망 매개변수에 대한 복원 계층을 사용하는 개념의 개략도를 도시한다. 도 6은 복원 계층 i, 예를 들어 제2 복원 계층, 복원 계층 i-1, 예를 들어 제1 복원 계층, 및 신경망(NN) 계층 p, 예를 들어 계층로, 도 3의 행렬(15a)와 같은 어레이 또는 행렬의 형태로 표현되는, 도 3의 계층(10b)를 도시한다.
- [0069] 도 6은 신경망을 정의하는 신경망 매개변수(13)를 복원하기 위한 장치(310)의 개념을 도시한다. 따라서, 장치는 이것은 예를 들어 연합 학습 프로세스 동안 이전에 전송도리 수 있고, 예를 들어, 제1 복원 계층에 대해, 예를 들어 복원 계층 i-1에 대해, 제 1 복원 계층 신경망 매개변수로 불릴 수 있는 제1 신경망 매개변수(13a)를 유도하도록 구성되어, 신경망 매개변수(13) 당, 제 2 복원 계층 신경망 매개변수 값을 산출한다. 따라서, 제1 및 제 2 복원 계층의 2개의 기여 값은 NN 매개변수당 획득될 수 있고 제1 및/또는 제2 NN 매개변수 값의 코딩/디코딩은 도 2 및 도 3에 따른 종속 양자화 및/또는 도 4 및 5에 설명된 바와 같은 양자화 인덱스의 산술 코딩/디코딩을 사용할 수 있다. 제2 신경망 매개변수(13b)는 제 1 표현 계층의 매개변수와 결합될 때, 신경 표현 측면에서 자체 포함된 의미를 갖지 않고, 신경망 표현, 즉 최종 신경망 매개변수로 이어질 수 있다.
- [0070] 게다가, 장치는 각 신경망 매개변수에 대해, 예를 들어 요소별 덧셈 및/또는 곱셈을 사용하여 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합(CB)함으로써, 신경망 매개변수(13)를 복원(314)하도록 구성된다.
- [0071] 추가로, 도 6은 제1 복원 계층에 대해, 예를 들어 신경망 매개변수(13)당 제1 복원 계층 신경망 매개변수 값을 포함하는 복원 계층(i-1)에 대해 제1 신경망 매개변수(13a)를 사용함으로써, 신경망을 정의하는, 신경망 매개변수(13)를 인코딩하기 위한 장치(320)에 대한 개념을 도시한다. 따라서, 장치는 제2 복원 계층의 경우, 예를 들어 복원 층 i에 대해, 제2 신경망 매개변수(13b)를 신경망 매개변수(13)당 제 2 복원 계층 신경망 매개변수 값을 포함하는 데이터 스트림으로 인코딩(122)하도록 구성되고, 여기서 신경망 매개변수(13)는, 각 신경망 매개변수에 대해, 예를 들어 요소별 덧셈 및/또는 곱셈을 사용하여 제 1 복원 계층 신경망 매개변수 값과 제 2 복원 계층 신경망 매개변수 값을 결합(CB)함으로써, 복원 가능하다.
- [0072] 선택적으로, 장치(310)는 제1 복원 계층에 대한 제1 신경망 매개변수를 데이터 스트림(14) 또는 별도의 데이터 스트림으로부터 디코딩(316)하도록 구성될 수 있다.
- [0073] 간단히 말해서, 신경망 매개변수(13)의 분해는 매개변수의 보다 효율적인 인코딩 및/또는 디코딩 및 전송을 가능하게 할 수 있다.
- [0074] 이하, 신경망 코딩 개념을 포함하는 추가 실시예가 개시된다. 이하 설명은 개별적으로 및 조합하여 위에서 설명된 실시예와 결합될 수 있는 추가 세부사항을 제공한다.
- [0075] 먼저, 본 발명의 실시예에 따른 종속 스칼라 양자화를 갖는 신경망의 매개변수의 엔트로피 코딩 방법이 제시될 것이다.
- [0076] 종속 스칼라 양자화를 사용하여 신경망 매개변수(13)의 세트(가중치, 가중치 매개변수 또는 매개변수라고도 함)의 매개변수 코딩 방법이 설명된다. 본 명세서에서 제시된 매개변수 코딩은 매개변수(13)의 종속 스칼라 양

자화(예를 들어, 도 3의 맥락에서 설명된 바와 같이) 및 획득된 양자화 인덱스(56)의 엔트로피 코딩(예를 들어, 도 5의 맥락에서 설명된 바와 같이)을 포함한다. 디코딩 측에서, 복원된 신경망 매개변수(13)의 세트는 양자화 인덱스(56)의 엔트로피 디코딩(예를 들어, 도 4의 맥락에서 설명된 바와 같이) 및 신경망 매개변수(13)의 의존적 복원(예를 들어, 도 2의 맥락에서 설명된 바와 같이)에 의해 획득된다. 독립 및 스칼라 양자화 및 엔트로피 코딩을 사용한 매개변수 코딩과 달리, 신경망 매개변수(13)에 대한 허용 가능한 복원 레벨 세트는 복원 순서에서 현재 신경망 매개변수(13')에 선행하는 전송된 양자화 인덱스(56)에 의존한다. 아래에 설명된 제시는 종속 스칼라 양자화에 사용되는 복원 레벨을 지정하는 양자화 인덱스의 엔트로피 코딩 방법을 추가로 설명한다.

[0077] 설명은 주로 신경망 압축에서 신경망 매개변수 계층의 손실 코딩을 대상으로 하지만, 손실 코딩의 다른 영역에도 적용될 수 있다.

[0078] 장치의 방법론은 다음으로 구성된, 여러 주요 부분으로 나뉠 수 있다.

[0079] 1. 양자화

[0080] 2. 무손실 인코딩

[0081] 3. 무손실 디코딩

[0082] 아래에 설명된 실시예의 주요 이점을 이해하기 위해, 먼저 신경망 주제와 매개변수 코딩 관련 방법에 대해 간략하게 소개한다. 그렇지만, 개시된 모든 측면, 특징 및 개념은 개별적으로 또는 본 명세서에서 설명된 실시예와 조합하여 사용될 수 있다.

[0083] 2 양자화 및 엔트로피 코딩에 대한 관련 방법

[0084] 멀티미디어 콘텐츠 설명 및 분석을 위한 신경망 압축을 위한 MPEG-7 파트 17 표준 작업 초안 2 [2]은 신경망 매개변수 코딩을 위해 독립 스칼라 양자화 및 엔트로피 코딩을 적용한다.

[0085] 2.1 스칼라 양자화기

[0086] 신경망 매개변수는 스칼라 양자화기를 사용하여 양자화된다. 양자화의 결과, 매개변수(13)에 대해 허용 가능한 값 세트는 감소된다. 즉, 신경망 매개변수는 소위 복원 레벨의 가산 세트(실제로는 유한 세트)에 매핑된다. 복원 레벨 세트는 가능한 신경망 매개변수 값 세트의 적절한 하위 세트를 나타낸다. 다음의 엔트로피 코딩을 단순화하기 위해, 허용 가능한 복원 레벨은 비트스트림(14)의 일부로 전송되는 양자화 인덱스(56)로 표시된다. 디코더 측에서, 양자화 인덱스(56)는 복원된 신경망 매개변수(13)에 매핑된다. 복원된 신경망 매개변수(13)에 대한 가능한 값은 복원 레벨의 세트(52)에 대응한다. 인코더 측에서 스칼라 양자화의 결과는 (정수) 양자화 인덱스(56) 세트이다.

[0087] 이 애플리케이션에서는 균일 복원 양자화기(URQ)가 사용된다. 기본 설계는 도 7에 나와 있다. 도 7은 균일 복원 양자화기의 예시를 도시한다. URQ는 복원 레벨의 간격이 동일하다는 특성이 있다. 두 개의 이웃하는 복원 레벨 사이의 거리 $\Delta(QP)$ 를 양자화 단계 크기라고 한다. 복원 레벨 중 하나는 0과 같다. 따라서 사용 가능한 복원 레벨의 전체 세트, 예를 들어, $\{t' \in \mathbb{N}_0\}$ 는 양자화 단계 크기 $\Delta(QP)$ 에 의해 고유하게 지정된다. 복원된 가중치 매개변수 $t'(13')$ 에 대한 양자화 인덱스 $q(56)$ 의 디코더 매핑은 원칙적으로 간단한 공식으로 제공된다:

[0088]
$$t' = q \cdot \Delta.$$

[0089] 이러한 맥락에서, "독립 스칼라 양자화"는 임의의 가중치 매개변수(13)에 대한 양자화 인덱스 $q(56)$ 이 주어지면, 연관된 복원된 가중치 매개변수 $t'(13')$ 은 다른 가중치 매개변수에 대한 모든 양자화 인덱스와 독립적으로 결정될 수 있다는, 속성을 갖는다.

[0090] 2.1.1 인코더 동작: 양자화

[0091] 신경망 압축 표준은 비트스트림 구문과 복원 프로세스만 지정한다. 원래 신경망 매개변수(13)의 주어진 세트와 주어진 양자화 단계 크기(QP)에 대한 매개변수 코딩을 고려하면, 인코더는 많은 자유도를 갖는다. 계층(10a, 10b)에 대한 양자화 인덱스 $q_k(56)$ 가 주어지면, 엔트로피 코딩은 데이터를 비트스트림(14)에 기록하기 위해(즉, 산술 코드워드를 구성하기 위해) 고유하게 정의된 알고리즘을 따라야 한다. 그러나 가중치 매개변수의 원본 세트(예: 계층)가 주어지면 양자화 인덱스 $q_k(56)$ 를 얻기 위한 인코더 알고리즘은 신경망 압축 표준의 범위를 벗어난다. 다음 설명을 위해, 각 신경망 매개변수(13)에 대한 양자화 단계 크기(QP)가 알려져 있다고 가정한다.

여전히 인코더는 각 신경망(가중치) 매개변수 $t_k(13)$ 에 대해 양자화기 인덱스 $q_k(56)$ 을 자유롭게 선택할 수 있다. 양자화 인덱스의 선택은 왜곡(또는 복원/근사 품질)과 비트 전송률 모두를 결정하기 때문에, 사용된 양자화 알고리즘은 생성된 비트스트림(14)의 윌-왜곡 성능에 상당한 영향을 미친다.

가장 간단한 양자화 방법은 신경망 매개변수 $t_k(13)$ 을 가장 가까운 복원 레벨로 반올림한다(가장 가까운 이웃 양자화라고도 함). 일반적으로 사용되는 URQ에 대해 해당 양자화 인덱스 $q_k(56)$ 는 다음에 따라 결정될 수 있다:

$$q_k = \text{sgn}(t_k) \cdot \left\lfloor \frac{|t_k|}{\Delta_k} + \frac{1}{2} \right\rfloor$$

여기서 $\text{sgn}()$ 은 부호 함수이고 연산자 $\lfloor \cdot \rfloor$ 는 인수보다 작거나 같은 가장 큰 정수를 반환한다. 이 양자화 방법은 MSE 왜곡

$$D = \sum_k D_k = \sum_k (t_k - q_k \cdot \Delta_k)^2$$

을 최소화하지만, 최종 매개변수 레벨(가중치 레벨) $q_k(56)$ 을 전송하는 데 필요한 비트 전송률은 완전히 무시한다. 이 방법은 MSE 왜곡 측정에 국한되지 않고, 또한 다른 모든 왜곡 측정, 예를 들어,

$$D^{MAE} = \sum_k D_k^{MAE} = \sum_k |t_k - q_k \cdot \Delta_k|$$

에 따른 MAE 왜곡이 사용될 수 있다는 것에 유의한다. 일반적으로, 반올림이 0으로 편향되면 더 나은 결과를 얻을 수 있다:

$$q_k = \text{sgn}(t_k) \cdot \left\lfloor \frac{|t_k|}{\Delta_k} + a \right\rfloor, \text{ 이 때 } 0 \leq a < \frac{1}{2}$$

양자화 과정이 라그랑지(Lagrangian) 함수 $D + \lambda \cdot R$ 를 최소화하면 윌-왜곡 감지에서 더 나은 결과를 얻을 수 있으며, 여기서 D 는 신경망 매개변수 세트의 왜곡(예: MSE 왜곡 또는 MAE 왜곡)을 나타내고, R 은 양자화 인덱스 $q_k(56)$ 를 전송하는 데 필요한 비트 수를 지정하고, λ 는 라그랑주 승수이다.

양자화 단계 크기가 주어지면, 라그랑주 승수 λ 와 양자화 단계 크기 사이의 다음 관계가 자주 사용된다:

$$\lambda = c_1 \cdot \Delta^2$$

여기에서 c_1 은 신경망 매개변수 세트에 대한 상수 인자를 나타낸다. 왜곡 및 비율의 라그랑주 함수 $D + \lambda \cdot R$ 를 최소화하는 것을 목표로 하는 양자화 알고리즘은 윌-왜곡 최적화 양자화(RDOQ)라고도 한다. MSE 또는 가중 MSE(또는 각각 MAE)를 사용하여 왜곡을 측정하면, 가중치 매개변수 세트(예: 계층)에 대한 양자화 인덱스 $q_k(56)$ 는 다음 비용 측정이 최소화되는 방식으로 결정되어야 한다:

$$D + \lambda \cdot R = \sum_k \alpha_k \cdot (t_k - \Delta_k \cdot q_k)^2 + \lambda \cdot R(q_k | q_{k-1}, q_{k-2}, \dots)$$

이때, 신경망 매개변수 인덱스 k 는 신경망 매개변수(13)의 코딩 순서(또는 스캐닝 순서)를 지정한다. 항 $R(q_k | q_{k-1}, q_{k-2}, \dots)$ 은 양자화 인덱스 $q_k(56)$ 를 전송하는데 필요한 비트 수(또는 그 추정치)를 나타낸다. 조건은 (결합 또는 조건부 확률의 사용으로 인해) 특정 양자화 인덱스 q_k 에 대한 비트 수는 일반적으로 코딩 순서로, 예를 들어 일반적인 순서(14')로, 이전 양자화 인덱스 q_{k-1} , q_{k-2} 등에 대해 선택된 값에 따라 다르다. 상기 수학식의 인자 α_k 는 개별 신경망 매개변수(13)의 기여도에 가중치를 부여하는 데 사용할 수 있다. 이하, 우리는 일반적으로 모든 가중치 인자 α_k 가 1과 같다고(그러나 알고리즘은 다른 가중치 요소를 고려할 수 있는 방식으로 간단하게 수정할 수 있음) 가정한다. 사실, 최근접 이웃 양자화는, 멀티미디어 콘텐츠 설명 및 분석을 위한 신경망 압축을 위한 MPEG-7 파트 17 표준의 작업 초안 2에 적용되는, $\lambda=0$ 인 사소한 경우이다.

- [0106] 2.2 엔트로피 코딩
- [0107] 이전 단계에서 적용된 균일 양자화의 결과, 가중치 매개변수는 소위 복원 레벨의 유한 세트에 매핑된다. 이들은 (정수) 양자화기 인덱스(56)(매개변수 레벨 또는 가중치 레벨이라고도 함) 및 예를 들어 전체 계층에 대해 고정될 수 있는 양자화 단계 크기(QP)로 나타낼 수 있다. 계층의 모든 양자화된 가중치 매개변수를 복원하기 위해, 계층의 단계 크기(QP) 및 치수는 디코더에 의해 알려져 있을 수 있다. 이들은 예를 들어 개별적으로 전송될 수 있다.
- [0108] 2.2.1 컨텍스트 적응 이진 산술 코딩(CABAC)을 사용한 양자화 인덱스 인코딩
- [0109] 양자화 인덱스(56)(정수 표현)는 엔트로피 코딩 기술을 사용하여 전송된다. 따라서 가중치 계층은 스캔을 사용하여 양자화된 가중치 레벨 시퀀스에 매핑된다. 예를 들어, 행 우선 스캔 순서는 행렬의 맨 위 행부터 시작하여 포함된 값을 왼쪽에서 오른쪽으로 인코딩하는 데 사용할 수 있다. 이러한 방식으로 모든 행은 위에서 아래로 인코딩된다. 스캔은 예를 들어, 뉴런 상호 연결(11)의 가중치와 관련될 수 있는 신경망 매개변수(13)를 포함하는 공통 순차 순서(14')를 따라, 행렬(15a)에 대해 도 3에 도시된 바와 같이 수행될 수 있다. 행렬은 가중치의 층, 예를 들어 층 $p-1(10a)$ 과 층 $p(10b)$ 사이의 가중치 또는 은닉 계층과 도 3 및 도 1에 각각 도시된 뉴런 상호접속(11)의 입력 층을 나타낼 수 있다. 다른 스캔을 적용할 수 있다. 예를 들어, 행렬(예를 들어, 도 2 또는 3의 행렬(15a))은 행 우선 스캔을 적용하기 전에, 전치되거나 수평 및/또는 수직으로 뒤집힐 수 있고/거나 왼쪽 또는 오른쪽으로 90/180/270도 회전될 수 있다.
- [0110] 도 3 및 5와 관련하여 설명된 바와 같이, 실시예에 따른 장치는 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈(84)에 대해 현재 신경망 매개변수(13')에 대한 상태에 의존하는 확률 모델을 사용하여 이진 산술 코딩을 사용하여 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 데이터 스트림(14)으로 인코딩하도록 구성될 수 있다. 확률 모델을 이용한 이진 산술 코딩은 CABAC(Context-Adaptive Binary Arithmetic Coding)일 수 있다.
- [0111] 즉, 실시예에 따라 레벨의 코딩을 위해 CABAC가 사용된다. 자세한 내용은 [3]을 참조한다. 따라서 양자화된 가중치 레벨 $q(56)$ 은 일련의 이진 기호 또는 구문 요소, 예를 들어 빈(이진 결정)으로 분해되고, 이는 다음에 CABAC(이진 산술 코더)로 전달될 수 있다.
- [0112] 제 1 단계에서, 이진 구문 요소 sig_flag 는 해당 레벨이 0과 같은지 여부를 지정하는 양자화된 가중치 레벨에 대해 유도된다. 다시 말해, 도 4에 도시된 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)가 0과 같거나 같지 않음을 나타내는 유의성 빈을 포함할 수 있다.
- [0113] sig_flag 가 1과 같으면 추가 이진 구문 요소인 $sign_flag$ 가 유도된다. 빈은 현재 가중치 레벨이 양수인지(예: $bin = 0$) 또는 음수(예: $bin = 1$)인지 나타낸다. 다시 말해, 도 4에 도시된 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)가 0보다 크거나 0보다 작은 것을 나타내는 부호 빈(86)을 포함할 수 있다.
- [0114] 다음에, 빈의 단항 시퀀스가 인코딩되고, 다음과 같이 고정 길이 시퀀스가 인코딩된다.
- [0115] 변수 k 는 음이 아닌 정수로 초기화되고 X 는 $1 \ll k$ 로 초기화된다.
- [0116] 양자화된 가중치 레벨의 절대값이 X 보다 크다는 것을 나타내는, 하나 이상의 구문 요소 $abs_level_greater_X$ 가 인코딩된다. $abs_level_greater_X$ 가 1이면, 변수 k 가 업데이트되고(예를 들어, 1 증가), 다음에 $1 \ll k$ 가 X 에 추가되고 추가 $abs_level_greater_X$ 가 인코딩된다. 이 절차는 $abs_level_greater_X$ 가 0이 될 때까지 계속된다. 그 후, 길이 k 의 고정 길이 코드는 양자화기 인덱스의 인코딩을 완료하기에 충분하다. 예를 들어, 변수 $rem = X - |q|$ 가 k 비트를 사용하여 인코딩될 수 있다. 또는 대안으로, 변수 rem' 은 k 비트를 사용하여 인코딩되는 $rem' = (1 \ll k) - rem - 1$ 로 정의될 수 있다. k 비트의 고정 길이 코드에 대한 변수 rem 의 임의의 다른 매핑은 대안적으로 사용될 수 있다.
- [0117] 다시 말해, 도 4에 도시된 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)의 절대값이 X 보다 크거나 크지 않은 것을 나타내는 X 보다 큰 빈을 포함할 수 있고, 여기서 X 는 0보다 큰 정수이다.
- [0118] 각 $abs_level_greater_X$ 이후에 k 를 1씩 증가시킬 때, 이 접근 방식은 인덱스 Golomb 코딩을 적용하는 것과 동하한다($sign_flag$ 가 고려되지 않는 경우).

- [0119] 추가적으로, 인코더 측과 디코더 측에서 최대 절대값 abs_max 를 알고 있다면, 전송되는 다음 $abs_level_greater_X$ 에 대해 $X \geq abs_max$ 가 유지될 때, $abs_level_greater_X$ 구문 요소의 인코딩이 종료될 수 있다.
- [0120] 2.2.2 컨텍스트 적응 이진 산술 코딩(CABAC)을 사용한 양자화 인덱스 디코딩
- [0121] 양자화된 가중치 레벨(56)의 디코딩(정수 표현)은 인코딩과 유사하게 작동한다. 디코더는 먼저 sig_flag 를 디코딩한다. 이것이 1이면, $sign_flag$ 및 $abs_level_greater_X$ 의 단항 시퀀스가 뒤따르며, 여기서 k 의 업데이트(따라서 X 의 증분)는 인코더에서와 동일한 규칙을 따라야 한다. 마지막으로, k 비트의 고정 길이 코드가 디코딩되어 정수로 해석된다(예를 들어, 둘 중 어느 것이 인코딩되었는지에 따라, rem 또는 rem'). 디코딩된 양자화된 가중치 레벨의 절대값 $|q|$ 은 X 로부터 복원되어, 고정 길이 부분을 형성할 수 있다. 예를 들어, rem 이 고정 길이 부분으로 사용된 경우 $|q|=X-rem$ 이다. 또는 rem' 이 인코딩된 경우, $|q|=X+1+rem'-(1 \leq k)$ 이다. 마지막 단계로, 부호는 디코딩된 $sign_flag$ 에 따라 $|q|$ 에 적용되어야 하므로, 양자화된 가중치 레벨 $q(56)$ 을 산출한다. 마지막으로, 양자화된 가중치 w 는 양자화된 가중치 레벨 q 에 스텝 크기 $\Delta(QP)$ 를 곱하여 복원된다.
- [0122] 다시 말해, 실시예에 따른 장치는 도 2 및 4와 관련하여 설명된 바와 같이, 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈(84)에 대해 현재 신경망 매개변수(13')에 대한 상태에 의존하는 확률 모델을 사용함으로써, 이진 산술 코딩을 사용하여 데이터 스트림(14)으로부터 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩하도록 구성될 수 있다.
- [0123] 도 5에 도시된 양자화 인덱스(56)의 이진화(82)의 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)가 0과 같거나 같지 않음을 나타내는 유의성 빈을 포함할 수 있다. 추가적으로 또는 대안적으로, 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)가 0보다 크거나 0보다 작은 것을 나타내는 부호 빈(86)을 포함할 수 있다. 또한, 적어도 하나의 빈은 현재 신경망 매개변수의 양자화 인덱스(56)의 절대값이 X 보다 크거나 크지 않음을 나타내는 X 보다 큰 빈을 포함할 수 있고, 여기서 X 는 0보다 큰 정수이다.
- [0124] 바람직한 실시예에서, k 는 0으로 초기화되고 다음과 같이 업데이트된다. 각 $abs_level_greater_X$ 가 1과 같으면 다음 규칙에 따라 필요한 k 업데이트가 수행된다. $X > X'$ 인 경우, k 는 1씩 증가하며 여기서 X' 는 애플리케이션에 따라 상수이다. 예를 들어 X' 는 인코더에 의해 유도되고 디코더에 신호되는 숫자(예: 0에서 100 사이)이다.
- [0125] 2.2.3 컨텍스트 모델링
- [0126] CABAC 엔트로피 코딩에서, 양자화된 가중치 레벨(56)에 대한 대부분의 구문 요소는 이진 확률 모델링을 사용하여 코딩된다. 각 이진 결정(빈)은 컨텍스트와 연결된다. 컨텍스트는 코딩된 빈 클래스에 대한 확률 모델을 나타낸다. 두 개의 가능한 빈 값 중 하나에 대한 확률은 해당 컨텍스트로 이미 코딩된 빈 값을 기반으로 각 컨텍스트에 대해 추정된다. 애플리케이션에 따라 다른 컨텍스트 모델링 접근 방식이 적용될 수 있다. 일반적으로 양자화된 가중치 코딩과 관련된 여러 빈의 경우, 코딩에 사용되는 컨텍스트는 이미 전송된 구문 요소를 기반으로 선택된다. 실제 애플리케이션에 따라, 예를 들어 SBMP 0 또는 HEVC 0 또는 VTM-4.0 0과 같은 다양한 확률 추정기가 선택될 수 있다. 선택은 예를 들어 압축 효율성과 복잡성에 영향을 미친다.
- [0127] 다시 말해, 도 5와 관련하여 설명된 확률 모델, 예. 컨텍스트(87)는 추가로 이전에 인코딩된 신경망 매개변수의 양자화 인덱스에 의존한다.
- [0128] 각각, 도 4와 관련하여 설명된 확률 모델, 예를 들어 컨텍스트(87)는 추가로 이전에 디코딩된 신경망 매개변수의 양자화 인덱스에 의존한다.
- [0129] 광범위한 신경망에 맞는 컨텍스트 모델링 방식은 다음과 같다. 가중치 행렬(계층)의 특정 위치(x, y)에서 양자화된 가중치 레벨 $q(56)$ 을 복호화하기 위해 현재 위치에 로컬 템플릿이 적용된다. 이 템플릿은 예를 들어 $(x-1, y)$, $(x, y-1)$, $(x-1, y-1)$ 등과 같은 여러 다른 (순서화된) 위치를 포함한다. 각 위치에 대해 상태 식별자가 유도된다.
- [0130] 바람직한 실시예(Si1로 표시됨)에서, 위치(x, y)에 대한 상태 식별자 $s_{x,y}$ 는 다음과 같이 유도된다: 위치 (x, y)가 행렬 외부로 가리키는 경우, 또는 위치 (x, y)에서 양자화된 가중치 레벨 $q_{x,y}$ 가 아직 디코딩되지 않았거나 0과 같으면, 상태 식별자 $s_{x,y}=0$ 이 된다. 그렇지 않으면, 상태 식별자는 $s_{x,y} = q_{x,y} < 0 ? 1 : 2$ 가 된다.
- [0131] 특정 템플릿의 경우, 상태 식별자의 시퀀스가 유도되고, 상태 식별자 값의 각각의 가능한 성상은 사용될 컨텍스트를 식별하는 컨텍스트 인덱스에 매핑된다. 템플릿 및 매핑은 상이한 구문 요소마다 다를 수 있다. 예를 들어

(정렬된) 위치 $(x-1, y)$, $(x, y-1)$, $(x-1, y-1)$ 을 포함하는 템플릿에서, 상태 식별자 $s_{x-1,y}$, $s_{x,y-1}$, $s_{x-1,y-1}$ 의 순서화된 시퀀스가 유도된다. 예를 들어, 이 시퀀스는 컨텍스트 인덱스 $C=s_{x-1,y}+3*s_{x,y-1}+9*s_{x-1,y-1}$ 에 매핑될 수 있다. 예를 들어, 컨텍스트 인덱스 C 는 `sig_flag`에 대한 컨텍스트의 수를 식별하는 데 사용될 수 있다.

[0132] 바람직한 실시예에서(접근법 1로 표시됨), 위치 (x,y) 에서 양자화된 가중치 레벨 $q_{x,y}$ 의 `sig_flag` 또는 `sign_flag`에 대한 로컬 템플릿은 단 하나의 위치 $(x-1, y)$ (즉, 왼쪽 이웃)로 구성된다. 연관된 상태 식별자 $s_{x-1,y}$ 는 바람직한 실시예 Si1에 따라 유도된다.

[0133] `sig_flag`의 경우 $s_{x-1,y}$ 값에 따라 세 가지 컨텍스트 중 하나가 선택되거나, `sign_flag`의 경우 $s_{x-1,y}$ 의 값에 따라 세 가지 다른 컨텍스트 중 하나가 선택된다.

[0134] 다른 바람직한 실시예에서(접근법 2로 표시됨), `sig` 플래그에 대한 로컬 템플릿에는 순서가 지정된 세 위치 $(x-1, y)$, $(x-2, y)$, $(x-3, y)$ 가 포함된다. 상태 식별자 $s_{x-1,y}$, $s_{x-2,y}$, $s_{x-3,y}$ 의 연관된 시퀀스는 바람직한 실시예 Si2에 따라 유도된다.

[0135] `sig_flag`의 경우, 컨텍스트 인덱스 C 는 다음과 같이 유도된다.

[0136] $s_{x-1,y} \neq 0$ 이면 $C=0$ 이다. 그렇지 않고 $s_{x-2,y} \neq 0$ 이면 $C=1$ 이다. 그렇지 않고 $s_{x-3,y} \neq 0$ 이면 $C=2$ 이다. 그렇지 않으면 $C=3$ 이다.

[0137] 이것은 또한 다음 방정식으로 표현될 수 있다.

$$C = (s_{x-1,y} \neq 0) ? 0 : ((s_{x-2,y} \neq 0) ? 1 : ((s_{x-3,y} \neq 0) ? 2 : 3))$$

[0139] 동일한 방법으로, 왼쪽에 있는 이웃의 수는 컨텍스트 인덱스 C 가 왼쪽으로 0이 아닌 다음 가중치까지의 거리와 같도록(템플릿 크기를 초과하지 않음) 증가 또는 감소될 수 있다.

[0140] 예를 들어, 각 `abs_level_greater_X` 플래그는 두 컨텍스트의 고유 세트를 적용할 수 있다. 그런 다음 두 컨텍스트 중 하나가 `sign_flag`의 값에 따라 선택된다.

[0141] 바람직한 실시예에서, 사전 정의된 숫자 X' 보다 적은 X 가 있는 `abs_level_greater_X` 플래그의 경우, 다른 컨텍스트는 X 및/또는 `sign_flag`의 값에 따라 구별된다.

[0142] 바람직한 실시예에서, X 가 사전 정의된 숫자 X' 보다 크거나 같은 `abs_level_greater_X` 플래그의 경우, 다른 컨텍스트는 X 에 따라 구분된다.

[0143] 또 다른 바람직한 실시예에서, X 가 사전 정의된 숫자 X' 보다 크거나 같은 `abs_level_greater_X` 플래그는 고정 코드 길이 1을 사용하여 인코딩된다(예: 산술 코더의 바이패스 모드 사용).

[0144] 더욱이, 구문 요소의 일부 또는 전부는 컨텍스트를 사용하지 않고 인코딩될 수도 있다. 대신 1비트의 고정 길이로 인코딩된다. 예를 들어 CABAC의 소위, 바이패스 빈을 사용한다.

[0145] 다른 바람직한 실시예에서, 고정 길이 나머지 `rem`은 바이패스 모드를 사용하여 인코딩된다.

[0146] 또 다른 바람직한 실시예에서, 인코더는 미리 정의된 숫자 X' 를 결정하고, 기호에 따라 $X < X'$ 인 각 구문 요소 `abs_level_greater_X`에 대해 두 개의 컨텍스트를 구별하고, $X \geq X'$ 인 각 `abs_level_greater_X`에 대해 하나의 컨텍스트를 사용한다.

[0147] 즉, 확률 모델, 예를 들어 도 5와 관련하여 설명된 바와 같이, 컨텍스트(87)는, 현재 신경망 매개변수가 관련된 부분에 인접하는 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 양자화 인덱스에 따라 확률 모델의 하위 세트에서 현재 신경망 매개 변수에 대해 선택될 수 있다(103).

[0148] 부분은 (정렬된) 위치 $(x-1, y)$, $(x, y-1)$, $(x-1, y-1)$ 을 포함하는 템플릿, 예를 들어 위에서 설명한 템플릿에 의해 정의될 수 있다.

[0149] 각각, 확률 모델은 도 5와 관련하여 설명된 바와 같이, 현재 신경망 매개변수가 관련된 부분에 인접하는 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 양자화 인덱스에 따라 확률 모델의 하위 세트에서 현재 신경망 매개 변수에 대해 선택될 수 있다.

- [0150] 3 추가 방법
- [0151] 이하는 복원된 계층, 예를 들어 도 6의 신경망 계층 p이 여러 하위 계층, 예를 들어, 개별적으로 전송될 수 있는 도 6의 복원 계층 i-1 및 복원 계층 i의 구성이 되는, 신경망(10)의 압축/전송을 위한 추가 및 이에 따른 선택적 방법을 설명한다.
- [0152] 3.1 기본 계층과 향상 계층의 개념
- [0153] 이 개념은 기본 계층과 향상 계층으로 표시된 두 가지 유형의 하위 계층을 소개한다. 그런 다음 복원 프로세스(예: 모든 하위 계층 추가)는 하위 계층에서 복원된 계층을 얻을 수 있는 방법을 정의한다. 기본 계층은 예를 들어 제 1 단계에서 효율적으로 표현되거나 압축/전송될 수 있도록 선택될 수 있는 기본 값을 포함한다. 향상 계층은 향상 정보, 예를 들어 (예를 들어, 원본 계층에 관한) 왜곡 측정을 줄이기 위해 (기본) 계층 값에 추가될 수 있는 차분 값을 포함한다. 다른 예에서, 기본 계층은 (작은 훈련 세트에 의한 훈련으로부터) 거친 값을 포함하고, 향상 계층은 개선 값을 포함한다(완전한 훈련 세트 또는 더 일반적으로 다른 훈련 세트를 기반으로 함). 하위 계층은 별도로 저장/전송될 수 있다.
- [0154] 바람직한 실시예에서, 압축될 계층 L_R , 예를 들어 신경망 매개변수, 예를 들어 도 2 및 3에서 행렬(15a)에 의해 표현될 수 있는 가중치와 같은 신경망 가중치의 계층은, 기본 계층 L_B 및 하나 이상의 향상 계층 $L_{E,1}, L_{E,2}, \dots, L_{E,N}$ 으로 분해된다. 그런 다음, 제 1 단계에서 기본 계층이 압축/전송되고 다음 단계에서 향상 계층 $L_{E,1}, L_{E,2}, \dots, L_{E,N}$ 이 (별도로) 압축/전송된다.
- [0155] 다른 바람직한 실시예에서, 복원된 계층 L_R 은 다음에 따라 모든 하위 계층 $L_{S,N}$ 을 (요소별로) 추가함으로써 획득될 수 있다:
- [0156]
$$L_R = \sum_{i=0}^N L_{S,N}$$
- [0157] 추가의 바람직한 실시예에서, 복원된 계층 L_R 은 다음에 따라 모든 하위 계층 $L_{S,N}$ 을 (요소별로) 곱함으로써 획득될 수 있다:
- [0158]
$$L_R = \prod_{i=0}^N L_{S,N}$$
- [0159] 즉, 본 발명에 따른 실시예는, 신경망 매개변수당, 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값의 매개변수 별 합 또는 매개변수 별 곱에 의해, 복원된 계층 L_R 의 형태로 또는 예를 들어 복원된 계층 L_R 을 사용하여, 신경망 매개변수(13)를 복원하도록 구성된, 장치를 포함한다.
- [0160] 각각, 실시예에 따른 신경망 매개변수(13)를 인코딩하기 위한 장치에 대해 신경망 매개변수(13)는 신경망 매개변수당 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값의 매개변수 별 합 또는 매개변수 별 곱에 의해 복원 가능하다. 추가의 바람직한 실시예에서, 2.1 및/또는 2.2의 방법은 하위 세트 또는 모든 하위 계층에 적용된다.
- [0161] 특히 바람직한 구현예에서, 컨텍스트 모델링을 사용하는 엔트로피 코딩 방식(예: 2.2.3과 유사)은 적용되지만 다음 규칙 중 하나 이상에 따라 하나 이상의 컨텍스트 모델 세트를 추가한다:
- [0162] a) 각 하위 계층은 고유한 컨텍스트 세트를 적용한다. 즉, 본 발명에 따른 실시예는, 데이터 스트림 또는 별도의 데이터 스트림으로부터 제1 복원 계층에 대한 제1 신경망 매개변수(13a)를 인코딩/디코딩하고, 제1 및 제2 복원 계층에 대한 개별 확률 컨텍스트를 사용하여 컨텍스트 적응 엔트로피 인코딩에 의해 데이터 스트림으로부터 제2 복원 계층에 대한 제2 신경망 매개변수(13b)를 인코딩/디코딩하도록 구성되는 장치를 포함한다.
- [0163] b) 인코딩될 향상 계층의 매개변수에 대해 선택된 컨텍스트 세트는 코딩 순서(예: 기본 계층)의 선행 계층에서 같은 위치에 있는 매개변수의 값에 따라 달라진다. 공동 배치된 매개변수가 0과 같을 때마다 컨텍스트 모델의 제 1 세트가 선택되고 그렇지 않으면 제 2 세트가 선택된다. 다시 말해, 본 발명에 따른 실시예는 1차 복원 계층 신경망 매개변수 값, 예를 들어 코딩 순서에서 선행 계층(예: 기본 계층)에서 같은 위치에 있는 매개변수의 값에 의존하는 확률 모델을 사용하는 상황 적응형 엔트로피 인코딩에 의해, 제 2 복원 계층 신경망 매개변수

값, 예를 들어, 항상 계층의 매개변수를 데이터 스트림으로 인코딩하도록 구성되는 장치를 포함한다. 추가 실시예에는 제 1 복원 계층 신경망 매개변수 값에 따라 확률 컨텍스트 세트의 집합 중 확률 컨텍스트 세트를 선택하고, 제1 복원 계층 신경망 매개변수 값에 따라 선택된 확률 컨텍스트 세트 중 사용될 확률 컨텍스트를 선택함으로써 컨텍스트 적응 엔트로피 인코딩을 통해 제 2 복원 계층 신경망 매개변수 값을 데이터 스트림으로 인코딩하도록 구성된 장치를 포함한다. 각각, 실시예에 따른 신경망 매개변수 디코딩 장치(13)에 대해, 상기 장치는 제1 복원 계층 신경망 매개변수 값에 의존하는 확률 모델을 사용하여 컨텍스트 적응 엔트로피 디코딩에 의해 데이터 스트림으로부터 제2 복원 계층 신경망 매개변수 값을 디코딩하도록 구성될 수 있다. 각각, 추가 실시예에는 제 1 복원 계층 신경망 매개변수 값에 따라 확률 컨텍스트 세트의 집합 중 확률 컨텍스트 세트를 선택하고, 제1 복원 계층 신경망 매개변수 값에 따라 선택된 확률 컨텍스트 세트 중 사용될 확률 컨텍스트를 선택함으로써, 컨텍스트 적응 엔트로피 디코딩에 의해 데이터 스트림으로부터 제2 복원 계층 신경망 매개변수 값을 디코딩하도록 구성된 장치를 포함한다.

[0164] c) 인코딩될 항상 계층의 매개변수에 대해 선택된 컨텍스트 세트는 코딩 순서(예: 기본 계층)의 선행 계층에서 같은 위치에 있는 매개변수의 값에 따라 달라진다. 공동 배치된 매개변수가 0보다 작으면(음수) 컨텍스트 모델의 제 1 세트가 선택되고, 공동 배치된 매개변수가 0보다 크면(양수) 제 2 세트가 선택되고 그렇지 않으면 제 3 세트가 선택된다. 다시 말해서, 본 발명에 따른 실시예는 장치, 예를 들어, 인코딩을 위한 장치를 포함하고, 여기서 확률 컨텍스트 세트의 컬렉션은 3개의 확률 컨텍스트 세트를 포함하고, 장치는 제1 복원 계층 신경망 매개변수 값이 음수인 경우 확률 컨텍스트 세트의 집합 중 제1 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하고, 제1 복원 계층 신경망 매개변수 값이 양수인 경우 확률 컨텍스트 세트의 집합 중 제2 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하고, 제1 복원 계층 신경망 매개변수 값이 0인 경우 확률 컨텍스트 세트의 집합 중에서 제3 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하도록 구성된다. 각각, 실시예에 따른 신경망 매개변수를 디코딩하기 위한 장치(13)에 대해, 확률 컨텍스트 세트의 컬렉션은 3개의 확률 컨텍스트 세트를 포함할 수 있으며, 상기 장치는 제 1 복원 계층 신경망 매개변수 값이 음수인 경우 확률 컨텍스트 세트의 집합 중 제 1 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하고, 제1 복원 계층 신경망 매개변수 값이 양수인 경우, 확률 컨텍스트 세트의 집합 중 제2 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하고, 제1 복원 계층 신경망 매개변수 값이 0인 경우 확률 컨텍스트 세트의 집합 중에서 제3 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하도록 구성될 수 있다.

[0165] d) 인코딩될 항상 계층의 매개변수에 대해 선택된 컨텍스트 세트는 코딩 순서(예: 기본 계층)에서 선행 계층의 같은 위치에 있는 매개변수의 값에 따라 달라진다. 공동 배치된 매개변수의 (절대) 값이 X (여기서 X 는 매개변수)보다 클 때마다 컨텍스트 모델의 제 1 세트가 선택되고, 그렇지 않으면 제 2 세트가 선택된다. 즉, 본 발명에 따른 실시예는 장치를 포함하고, 여기서 확률 컨텍스트 세트의 컬렉션은 2개의 확률 컨텍스트 세트를 포함하고, 장치는 제 1 복원 계층 신경망 매개변수 값, 예를 들어 코딩 순서에서 선행 계층(예: 기본 계층)에서 같은 위치에 있는 매개변수의 값이 미리 결정된 값, 예를 들어, x 보다 큰 경우, 확률 컨텍스트 세트의 집합 중 제1 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로서 선택하고, 제1 복원 계층 신경망 매개변수 값이 미리 결정된 값보다 크지 않은 경우, 확률 컨텍스트 세트의 집합 중에서 선택된 확률 컨텍스트 세트로서 제2 확률 컨텍스트 세트를 선택하고, 제1 복원 계층 신경망 매개변수 값의 절대값이 미리 정해진 값보다 큰 경우 선택된 확률 컨텍스트 세트로서 확률 컨텍스트 세트의 컬렉션으로부터 제 1 확률 컨텍스트 세트를 선택하고, 제1 복원 계층 신경망 매개변수 값의 절대값이 미리 결정된 값보다 크지 않은 경우 확률 컨텍스트 세트의 집합 중에서 선택된 제2 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하도록 구성된다. 각각, 실시예에 따른 신경망 매개변수 디코딩 장치(13)에 대해, 확률 컨텍스트의 컬렉션은 2개의 확률 컨텍스트 세트를 포함할 수 있으며, 장치들은 제 1 복원 계층 신경망 매개변수 값이 미리 결정된 값, 예를 들어 x 보다 큰 경우, 확률 컨텍스트 세트들의 세트로부터 제1 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로서 선택하고, 제1 복원 계층 신경망 매개변수 값이 미리 결정된 값보다 크지 않은 경우 확률 컨텍스트 세트의 집합 중에서 선택된 확률 컨텍스트 세트로서 제2 확률 컨텍스트 세트를 선택하고, 제1 복원 계층 신경망 매개변수 값의 절대값이 미리 정해진 값보다 큰 경우 선택된 확률 컨텍스트 세트로서 확률 컨텍스트 세트의 컬렉션으로부터 제 1 확률 컨텍스트 세트를 선택하고, 제1 복원 계층 신경망 매개변수 값의 절대값이 미리 결정된 값보다 크지 않은 경우 확률 컨텍스트 세트의 집합 중에서 선택된 제2 확률 컨텍스트 세트를 선택된 확률 컨텍스트 세트로 선택하도록 구성된다.

[0166] 4 종속 스칼라 양자화를 사용한 신경망 매개변수 코딩

[0167] 이 섹션에서, 도 2 내지 도 4의 맥락에서 설명된 바와 같이, 본 발명에 따른 개념 및 실시예에 대한 추가의 선택적인 측면 및 특징이 개시된다.

- [0168] 다음은 신경망 매개변수 코딩에 대한 수정된 개념을 설명한다. 이전에 설명된 신경망 매개변수 코딩과 관련된 주요 변경은 신경망 매개변수(13)가 독립적으로 양자화 및 복원되지 않는다는 것이다. 대신에, 신경망 매개변수(13)에 대한 허용 가능한 복원 레벨은 복원 순서에서 선행하는 신경망 매개변수에 대해 선택된 양자화 인덱스(56)에 의존한다. 종속 스칼라 양자화의 개념은 수정된 엔트로피 코딩과 결합되며, 여기서 신경망 매개변수에 대한 확률 모델 선택(또는 대안적으로 코드워드 표 선택)은 허용 가능한 복원 레벨 세트에 의존한다. 그러나, 이전에 설명된 실시예는 다음에서 설명되는 임의의 특징에 의해 개별적으로 또는 조합하여 사용 및/또는 통합 및/또는 확장될 수 있다는 점에 유의해야 한다.
- [0169] 4.1 관련 신경망 매개변수 코딩에 비해 장점
- [0170] 신경망 매개변수의 종속 양자화의 이점은 허용 가능한 복원 벡터가 N차원 신호 공간에서 더 조밀하게 패키징된다는 것이다(여기서 N은 처리될 샘플 세트, 예를 들어 계층(10a, 10b)에서 샘플 또는 신경망 매개변수(13)의 수를 나타냄). 신경망 매개변수 세트에 대한 복원 벡터는 신경망 매개변수 세트의 순서화된 복원된 신경망 매개변수(또는 대안적으로 순서화된 복원된 샘플)를 참조한다. 종속 스칼라 양자화의 효과는 두 개의 신경망 매개변수의 가장 간단한 경우에 대해 도 8에 나와 있다. 도 8은 두 개의 가중치 매개변수의 간단한 경우에 대해 허용 가능한 복원 벡터 위치의 예를 보여준다: 도 8(a)는 독립 스칼라 양자화의 예를 도시하고; 도 8(b)는 종속 스칼라 양자화의 예를 도시한다. 도 8a는 독립 스칼라 양자화를 위한 허용 가능한 복원 벡터(201)(2d 평면의 포인트를 나타냄)를 보여준다. 알 수 있는 바와 같이, 제 2 신경망 매개변수 θ_2 (13)에 대한 허용 가능한 값 세트는 복원된 제 1 신경망 매개변수 θ_1 (13)에 대해 선택된 값에 의존하지 않는다. 도 8(b)는 종속 스칼라 양자화의 예를 보여준다. 독립 스칼라 양자화와 대조적으로, 제 2 신경망 매개변수 θ_2 (13)에 대한 선택 가능한 복원 값은 제 1 신경망 매개변수 θ_1 (13)에 대해 선택된 복원 레벨에 따라 다르다. 도 8b의 예에서, 제2 신경망 매개변수 θ_2 (13)에 대한 이용가능한 복원 레벨의 2개의 상이한 세트(52)가 있다(상이한 색상으로 예시됨). 제1 신경망 매개변수 θ_1 (13)에 대한 양자화 인덱스(56)가 짝수($\dots, -2, 0, 2, \dots$)이면, 제1 세트(청색 점)의 임의의 복원 레벨(201a)은 제2 신경망 매개변수 θ_2 (13)에 대해 선택될 수 있다. 그리고 제1 신경망 매개변수 θ_1 에 대한 양자화 인덱스(56)가 홀수($\dots, -3, -1, 1, 3, \dots$)이면, 제2 세트(적색 점)의 임의의 복원 레벨(201b)은 제2 신경망 매개변수 θ_2 (13)에 대해 선택될 수 있다. 예에서, 제 1 및 제 2 세트에 대한 복원 레벨은 양자화 단계 크기의 절반만큼 시프트된다(제 2 세트의 복원 레벨은 제 1 세트의 두 복원 레벨 사이에 위치한다).
- [0171] 신경망 매개변수(13)의 종속 스칼라 양자화는 N차원 단위 부피당 복원 벡터(201)의 주어진 평균 수에 대해, 신경망 매개변수(13)의 주어진 입력 벡터와 가장 가까운 이용 가능한 복원 벡터 사이의 거리의 기대값이 감소된다는 효과가 있다. 결과적으로, 신경망 매개변수의 입력 벡터와 복원된 신경망 매개변수 사이의 평균 왜곡은 주어진 평균 비트 수에 대해 감소될 수 있다. 벡터 양자화에서 이 효과를 공간 채우기 이득이라고 한다. 신경망 매개변수(13)의 세트에 대한 종속 스칼라 양자화를 사용하여, 고차원 벡터 양자화를 위한 잠재적인 공간 채우기 이득의 주요 부분이 이용될 수 있다. 그리고 벡터 양자화와 대조적으로, 복원 프로세스(또는 디코딩 프로세스)의 구현 복잡성은 독립 스칼라 양자화기를 사용한 관련 신경망 매개변수 코딩의 복잡성과 비슷하다.
- [0172] 4.2 개요
- [0173] 주요 변경 사항은 앞서 언급한 바와 같이 종속 양자화이다. 복원 순서 인덱스 $k > 0$ 을 갖는 복원된 신경망 매개변수 θ_k (13)는 연관된 양자화 인덱스 q_k (56)만이 아니라 복원 순서로 선행하는 신경망 매개변수에 대한 양자화 인덱스 $q_0, q_1, \dots, q_{k-1}$ 에도 의존한다. 종속 양자화에서 신경망 매개변수(13)의 복원 순서는 고유하게 정의되어야 한다. 전체 신경망 코덱의 성능은 양자화 인덱스 q_k (56)와 관련된 복원 레벨 세트에 대한 지식이 엔트로피 코딩에서도 활용되는 경우, 일반적으로 향상될 수 있다. 즉, 일반적으로 신경망 매개변수에 적용되는 복원 레벨 세트를 기반으로 컨텍스트(확률 모델) 또는 코드워드 표를 전환하는 것이 좋다.
- [0174] 엔트로피 코딩은 일반적으로 주어진 엔트로피 디코딩 프로세스에서 고유하게 지정된다. 그러나 관련 신경망 매개변수 코딩과 유사하게 원래의 신경망 매개변수가 주어지면 양자화 인덱스를 선택하는 데 많은 자유가 있다.
- [0175] 본 명세서에서 설명된 실시예는 계층별 신경망 코딩으로 제한되지 않는다. 또한 신경망 매개변수(13)의 유한 집

합의 신경망 매개변수 코딩에도 적용될 수 있다.

[0176] 특히, 이 방법은 단락 3.1에서 설명된 바와 같이 하위 계층에도 적용될 수 있다.

[0177] 4.3 신경망 매개변수의 종속 양자화

[0178] 신경망 매개변수의 종속 양자화(13)는, 신경망 매개변수(13)에 대한 이용 가능한 복원 레벨 세트가 (동일한 신경망 매개변수 세트, 예를 들어, 계층 또는 하위 계층의 내부에서) 복원 순서에서 선행하는 신경망 매개변수에 대해 선택된 양자화 인덱스에 의존하는, 개념을 말한다.

[0179] 바람직한 실시예에서, 여러 세트의 복원 레벨이 사전 정의되고 코딩 순서에서 이전 신경망 매개변수에 대한 양자화 인덱스에 기초하여, 현재 신경망 매개변수를 복원하기 위해 미리 정의된 세트 중 하나가 선택된다. 다시 말해, 실시예에 따른 장치는 현재 신경망 매개변수(13)에 대해, 이전, 예를 들어, 선행하는 신경망 매개변수에 대한 양자화 인덱스(58)에 따라, 복수(50)의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 선택(54)하도록 구성될 수 있다.

[0180] 복원 레벨의 세트를 정의하기 위한 바람직한 실시예는 단락 4.3. 1에서 설명된다. 선택한 복원 레벨의 식별 및 신호는 4.3.2절에 설명되어 있다. 단락 4.3.3은 현재 신경망 매개변수에 대해 미리 정의된 복원 레벨 세트 중 하나를 선택하기 위한 바람직한 실시예를 설명한다(복원 순서에서 이전 신경망 매개변수에 대해 선택된 양자화 인덱스를 기반으로 함).

[0181] 4.3.1 복원 레벨 세트

[0182] 바람직한 실시예에서, 현재 신경망 매개변수에 대한 허용 가능한 복원 레벨 세트는 (코딩 순서에서 이전 신경망 매개변수에 대한 양자화 인덱스를 기반으로 하여) 복원 레벨의 미리 정의된 세트(52)의 집합 중에서 선택된다 (2개 이상의 세트, 예를 들어 도 2 및 3의 세트 0 및 세트 1).

[0183] 바람직한 실시예에서, 매개변수는 양자화 단계 크기 $\Delta(QP)$ 를 결정하고, 모든 복원 레벨(모든 복원 레벨 세트에서)은 양자화 단계 크기 Δ 의 정수배를 나타낸다. 그러나 복원 레벨의 각 세트는 양자화 단계 크기 $\Delta(QP)$ 의 정수 배수의 하위 세트만을 포함한다는 점에 유의한다. 모든 복원 레벨 세트에 대한 모든 가능한 복원 레벨은 양자화 단계 크기 $\Delta(QP)$ 의 정수 배수를 나타내는, 종속양자화를 위한 이러한 구성은, 단일 복원 양자화기(URQ)의 확장으로 간주될 수 있다. 그것의 기본적인 이점은 복원된 신경망 매개변수(13)가 매우 낮은 계산 복잡성을 갖는 알고리즘에 의해 계산될 수 있다는 것이다(아래에서 더 자세히 설명됨).

[0184] 복원 레벨 세트는 완전히 분리될 수 있지만: 하나 이상의 복원 레벨이 여러 세트에 포함될 수도 있다(세트는 여전히 다른 복원 레벨이 다르다).

[0185] 바람직한 실시예에서, 신경망 매개변수에 대한 종속 스칼라 양자화는 정확히 두 개의 서로 다른 복원 레벨 세트를 사용한다. 0을 설정하고 1을 설정한다. 그리고 특히 바람직한 구현예에서, 신경망 매개변수 $t_k(13)$ 에 대한 두 세트의 모든 복원 레벨은 이 신경망 매개변수(13)에 대한 양자화 단계 크기 $\Delta_k(QP)$ 의 정수배를 나타낸다. 양자화 단계 크기 $\Delta_k(QP)$ 는 두 세트 모두에서 허용 가능한 복원 값에 대한 스케일링 인자를 나타낸다. 모든 신경망 매개변수(13)에 대해 동일한 두 세트의 복원 레벨이 사용된다.

[0186] 도 9에서, 복원 레벨의 2개 세트(세트 0 및 세트 1)에 대한 3개의 선호되는 구성((a)-(c))이 도시되어 있다. 도 9는 단일 양자화 단계 크기 $\Delta(QP)$ 에 의해 완전히 결정되는 두 세트의 복원 레벨을 사용한 종속 양자화의 예를 보여준다. 사용 가능한 두 가지 복원 레벨 세트는 서로 다른 색상으로 강조 표시된다(세트 0의 경우 파란색, 세트 1의 경우 빨간색). 세트 내부의 복원 레벨을 나타내는 양자화 인덱스의 예는 원 아래의 숫자로 표시된다. 속이 빈 원과 채워진 원은 복원 레벨 세트 내부의 두 가지 다른 하위 세트를 나타내고; 하위 세트는 복원 순서에서 다음 신경망 매개변수에 대한 복원 레벨 세트를 결정하는 데 사용될 수 있다. 도면은 두 가지 복원 레벨 세트가 있는 세 가지 기본 구성을 보여준다: (a) 두 세트는 0에 대해 분리되고 대칭이다; (b) 두 세트 모두 0과 동일한 복원 레벨을 포함하지만 그렇지 않으면 분리되고; 세트는 0을 중심으로 비대칭이다; (c) 두 세트 모두 0과 동일한 복원 레벨을 포함하지만 그렇지 않으면 분리된다. 두 세트 모두 0을 중심으로 대칭이다. 모든 복원 레벨은 양자화 단계 크기 Δ 의 정수 배수(IV)로 주어진 그리드에 있다는 점에 유의한다. 또한 특정 복원 레벨이 두 세트에 모두 포함될 수 있다는 점에 유의해야 한다.

[0187] 도 9(a)에 표시된 두 세트는 분리되어 있다. 양자화 단계 크기 $\Delta(QP)$ 의 각 정수 배수는 세트 중 하나에만 포함된다. 제 1 세트(세트 0)는 양자화 단계 크기의 모든 짝수 정수 배수(IV)를 포함하는 반면, 제 2 세트(세트 1)

는 양자화 단계 크기의 모든 홀수 정수 배수를 포함한다. 두 세트 모두에서 인접한 두 복원 레벨 사이의 거리는 양자화 단계 크기의 2배이다. 이 두 세트는 일반적으로 고속 양자화, 즉 신경망 매개변수의 분산이 양자화 단계 크기(QP)보다 훨씬 큰 설정에 적합하다. 그러나 신경망 매개변수 코딩에서 양자화기는 일반적으로 낮은 비율 범위에서 작동한다. 일반적으로, 많은 원본 신경망 매개변수(13)의 절대값은 양자화 단계 크기(QP)의 0이 아닌 배수보다 0에 더 가깝다. 그 경우, 양자화 세트(복원 레벨 세트) 모두에 0이 포함되는 것이 일반적으로 바람직하다.

[0188] 도 9(b)에 표시된 두 개의 양자화 세트는 모두 0을 포함한다. 세트 0에서, 0과 동일한 복원 레벨과 0보다 큰 제 1 복원 레벨 사이의 거리는 양자화 단계 크기(QP)와 동일한 반면, 2개의 이웃하는 복원 레벨 사이의 다른 모든 거리는 양자화 단계 크기의 2배와 같다. 마찬가지로 세트 1에서, 0과 동일한 복원 레벨과 0보다 작은 제1 복원 레벨 사이의 거리는 양자화 단계 크기와 동일한 반면, 2개의 이웃하는 복원 레벨 사이의 다른 모든 거리는 양자화 단계 크기의 2배와 같다. 두 복원 세트 모두 0을 중심으로 비대칭이다. 이것은 부호의 확률을 정확하게 추정하기 어렵게 하기 때문에, 비효율을 초래할 수 있다.

[0189] 두 세트의 복원 레벨에 대한 기본 구성이 도 9(c)에 나와 있다. 제 1 양자화 세트에 포함된 복원 레벨(도면에서 세트 0으로 표시됨)은 양자화 단계 크기의 짝수 배수를 나타낸다(이 세트는 실제로 도 9(a)의 세트 0과 동일하다). 제 2 양자화 세트(도면에서 세트 1로 표시됨)에는 양자화 단계 크기의 모든 홀수 정수배와 0과 같은 복원 레벨이 포함된다. 두 복원 세트는 모두 0에 대해 대칭이다. 0과 같은 복원 레벨은 두 복원 세트에 모두 포함되며, 그렇지 않으면 복원 세트가 분리된다. 두 복원 세트의 합세트는 양자화 단계 크기의 모든 정수 배수를 포함한다.

[0190] 다시 말해서, 예를 들어 신경망 매개변수(13)를 인코딩/디코딩하기 위한 장치를 포함하는 실시예에 따르면, 복수의 복원 레벨 세트(52)(52)의 복원 레벨 세트(52)의 수는 2이고(예를 들어, 세트 0, 세트 1), 복수의 복원 레벨 세트는 미리 결정된 양자화 단계 크기의 0 및 짝수 배수를 포함하는 제1 복원 레벨 세트(세트 0), 및 미리 결정된 양자화 단계 크기의 0 및 홀수 배수를 포함하는 제2 복원 레벨 세트(세트 1)를 포함한다.

[0191] 또한, 모든 복원 레벨 세트의 모든 복원 레벨은 미리 결정된 양자화 단계 크기(QP)의 정수 배수(IV)를 나타낼 수 있으며, 예를 들어 실시예에 따라, 신경망 매개변수(13)를 디코딩하기 위한 장치는 각각의 신경망 매개변수에 대해 설정된 선택된 복원 레벨 및 각각의 신경망 매개변수(13')에 대한 엔트로피 디코딩된 양자화 인덱스(58)에 따라 각 신경망 매개변수에 대해 중간 정수 값, 예를 들어 정수배(IV)를 유도하고, 각각의 신경망 매개변수(13)에 대해, 각각의 신경망 매개변수에 대한 중간값에 각각의 신경망 매개변수(13)에 대한 미리 결정된 양자화 단계 크기를 곱함으로써, 신경망 매개변수(13)을 역양자화하도록 구성될 수 있다.

[0192] 각각, 모든 복원 레벨 세트의 모든 복원 레벨은 미리 결정된 양자화 단계 크기(QP)의 정수 배수(IV)를 나타낼 수 있으며, 및 예를 들어 실시예에 따라, 신경망 매개변수(13)를 인코딩하기 위한 장치는 각각의 신경망 매개변수에 대해, 각각의 신경망 매개변수에 대해 설정된 선택된 복원 레벨 및 각각의 신경망 매개변수에 대한 엔트로피 인코딩된 양자화 인덱스에 따른 중간 정수 값을 유도하고, 각각의 신경망 매개변수에 대해, 각각의 신경망 매개변수에 대한 미리 결정된 양자화 단계 크기와 각각의 신경망 매개변수에 대한 중간 값을 곱함으로써, 이것이 역양자화 가능한 방식으로 신경망 매개변수를 양자화하도록 구성될 수 있다.

[0193] 본 명세서에서 설명된 실시예는 도 9에 도시된 구성으로 제한되지 않는다. 복원 레벨의 임의의 다른 2개의 상이한 세트가 사용될 수 있다. 다중 복원 레벨이 두 세트에 모두 포함될 수 있다. 또는 두 양자화 세트의 합집합이 양자화 단계 크기의 모든 가능한 정수 배수를 포함하지 않을 수 있다. 또한, 신경망 매개변수의 종속 스칼라 양자화를 위해 두 개 이상의 복원 레벨 세트를 사용할 수 있다.

[0194] 4.3.2 선택한 복원 레벨의 신호

[0195] 인코더가 허용 가능한 복원 레벨 중에서 선택하는 복원 레벨은 비트스트림(14) 내부에 표시되어야 한다. 종래의 독립 스칼라 양자화에서와 같이, 이것은 가중치 레벨로도 지칭되는 소위 양자화 인덱스(56)를 사용하여 달성될 수 있다. 양자화 인덱스(56)(또는 가중치 레벨)은 양자화 세트(52) 내부(즉, 복원 레벨 세트 내부)의 이용가능한 복원 레벨을 고유하게 식별하는 정수이다. 양자화 인덱스(56)는 (엔트로피 코딩 기술을 사용하여) 비트스트림(14)의 일부로 디코더에 전송된다. 디코더 측에서, 복원된 신경망 매개변수(13)는 (코딩/복원 순서에서 이전 양자화 인덱스에 의해 결정되는) 복원 레벨의 현재 세트(48) 및 현재 신경망 매개변수(13')에 대한 전송된 양자화 인덱스(56)에 기초하여, 고유하게 계산될 수 있다.

[0196] 바람직한 실시예에서, 복원 레벨 세트(또는 양자화 세트) 내의 복원 레벨에 대한 양자화 인덱스(56)의 할당은

다음 규칙을 따른다. 예를 들어, 도 9의 복원 레벨은 관련 양자화 인덱스(56)로 레이블링된다 (양자화 인덱스는 복원 레벨을 나타내는 원 아래의 숫자로 제공된다). 복원 레벨 세트가 0과 동일한 복원 레벨을 포함하는 경우, 0과 동일한 양자화 인덱스는 0과 동일한 복원 레벨에 할당된다. 1과 동일한 양자화 인덱스는 0보다 큰 최소 복원 레벨에 할당되고, 2와 동일한 양자화 인덱스는 0보다 큰 다음 복원 레벨에 할당되는(즉, 0보다 큰 제 2로 작은 복원 레벨) 등이다. 또는, 다시 말해서, 0보다 큰 복원 레벨은 값의 오름차순으로 0보다 큰 정수(즉, 1, 2, 3 등)로 레이블링된다. 유사하게, 양자화 인덱스 -1은 0보다 작은 가장 큰 복원 레벨에 할당되고, 양자화 인덱스 -2는 0보다 작은 다음(즉, 제 2로 큰) 복원 레벨에 할당되는 식이다. 즉, 0보다 작은 복원 레벨은 값의 내림차순으로 0보다 작은 정수(즉, -1, -2, -3 등)로 레이블링된다. 도 9의 예에서는, 설명된 양자화 인덱스 할당은 도 9(a)의 세트 1을 제외한 모든 양자화 세트에 대해 설명된다(0과 동일한 복원 레벨은 포함하지 않음).

[0197] 0과 같은 복원 레벨을 포함하지 않는 양자화 세트의 경우, 양자화 인덱스(56)를 복원 레벨에 할당하는 한 가지 방법은 다음과 같다. 0보다 큰 모든 복원 레벨은 0보다 큰 양자화 인덱스로 레이블이 지정되고(값의 오름차순), 0보다 작은 모든 복원 레벨은 0보다 작은 양자화 인덱스로 레이블링된다(값의 내림차순). 따라서, 양자화 인덱스(56)의 할당은 기본적으로 0과 동일한 복원 레벨을 포함하는 양자화 세트의 경우와 동일한 개념을 따른다. 0과 동일한 양자화 인덱스가 없다는 차이점이 있다(도 9(a)의 양자화 세트 1에 대한 레이블 참조). 그 측면은 양자화 인덱스(56)의 엔트로피 코딩에서 고려되어야 한다. 예를 들어, 양자화 인덱스(56)는 종종 절대값(0에서 최대 지원 값까지)을 코딩하여, 0이 아닌 절대값에 대해서는, 양자화 인덱스(56)의 부호를 추가로 코딩하여 전송된다. 0과 같은 양자화 인덱스(56)가 이용가능하지 않다면, 엔트로피 코딩은 절대 레벨에서 1을 뺀 값이 전송되는 방식으로 수정될 수 있고(해당 구문 요소의 값은 0에서 최대 지원 값 범위). 부호는 항상 전송된다. 대안으로서, 양자화 인덱스(56)를 복원 레벨에 할당하기 위한 할당 규칙이 수정될 수 있다. 예를 들어, 0에 가까운 복원 레벨 중 하나는 0과 동일한 양자화 인덱스로 레이블될 수 있다. 그런 다음 나머지 복원 레벨에는 다음 규칙에 따라 레이블이 지정된다. 0보다 큰 양자화 인덱스는 양자화 인덱스가 0인 복원 레벨보다 큰 복원 레벨에 할당된다(양자화 인덱스는 복원 레벨의 값에 따라 증가한다). 그리고 0보다 작은 양자화 인덱스는 양자화 인덱스가 0인 복원 레벨보다 작은 복원 레벨에 할당된다(양자화 인덱스는 복원 레벨의 값에 따라 감소한다). 그러한 할당에 대한 한 가지 가능성은 도 9(a)에서 괄호 안의 숫자로 설명된다(괄호 안에 숫자가 없으면, 다른 숫자가 적용된다).

[0198] 위에서 언급한 바와 같이, 바람직한 실시예에서, 두 개의 서로 다른 복원 레벨 세트(양자화 세트라고도 함)가 사용되고, 두 세트 내의 복원 레벨은 양자화 단계 크기(QP)의 정수 배수를 나타낸다. 여기에는 양자화 단계 크기가 (예를 들어, 비트스트림(14) 내부에 계층 양자화 매개변수를 전송함으로써) 계층 기반으로 또는 (예를 들어, 비트스트림(14) 내부에서 블록 양자화 매개변수를 전송함으로써) 신경망 매개변수(13)의 다른 유한 세트(예: 블록) 기반으로 수정된다.

[0199] 양자화 단계 크기(QP)의 정수 배수를 나타내는 복원 레벨의 사용은 디코더 측에서 신경망 매개변수(13)의 복원을 위해 계산적으로 낮은 복잡한 알고리즘을 가능하게 한다. 이것은 다음 도 9(c)의 바람직한 예를 기반으로 설명된다(다른 구성, 특히 도 9(a) 및 도 9(b)에 표시된 설정에도 유사한 간단한 알고리즘이 있다). 도 9(c)에 도시된 구성에서, 제 1 양자화 세트는 양자화 단계 크기(QP)의 모든 짝수 정수배를 포함하고, 제2 양자화 세트는 양자화 단계 크기의 모든 홀수 배수에 0과 동일한 복원 레벨(양자화 세트 모두에 포함됨)을 더한 것을 포함한다. 신경망 매개변수에 대한 복원 프로세스는 도 10의 의사 코드에 지정된 알고리즘과 유사하게 구현될 수 있다. 도 10은 신경망 매개변수(13)에 대한 복원 프로세스의 바람직한 예를 나타내는 의사 코드의 예를 보여 준다. k 는 현재 신경망 매개변수(13')의 복원 순서를 지정하는 인덱스를 나타내며, 현재 신경망 매개변수에 대한 양자화 인덱스(56)는 레벨[k](210)로 표시되며, 현재 신경망 매개변수(13')에 적용되는 양자화 단계 크기 Δ_k (QP)는 `quant_step_size[k]`로 표시되며, `trec[k]`(220)는 복원된 신경망 매개변수 t_k 의 값을 나타낸다. 변수 `setId[k]`(240)는 현재 신경망 매개변수(13')에 적용되는 복원 레벨 세트를 지정한다. 이것은 복원 순서에 따라 선행 신경망 매개변수를 기반으로 결정된다; `setId[k]`의 가능한 값은 0과 1이다. 변수 n 은 양자화 단계 크기(QP)의 정수 인수, 예를 들어, 중간 값 IV를 지정하고; 이것은 복원 레벨의 선택된 세트(즉, `setId[k]`의 값) 및 전송된 양자화 인덱스 레벨[k]에 의해 제공된다.

[0200] 도 10의 의사 코드에서, `level[k]`는 신경망 매개변수 t_k (13)에 대해 전송되는 양자화 인덱스(56)를 나타내고, `setId[k]`(0 또는 1과 같음)는 현재 복원 레벨 세트의 식별자를 지정한다(이것은 아래에서 더 상세히 설명되는 바와 같이 복원 순서로 선행 양자화 인덱스(56)에 기초하여 결정된다). 변수 n 은 양자화 인덱스 레벨[k] 및 세트 식별자 `setId[k]`에 의해 주어진 양자화 단계 크기(QP)의 정수배를 나타낸다. 신경망 매개변수(13)가 양자화

단계 크기 $\Delta_k(QP)$ 의 짝수 정수 배수를 포함하는 복원 레벨의 제 1 세트($setId[k] == 0$)를 사용하여 코딩되면, 변수 n 은 전송된 양자화 인덱스(56)의 2배이다. 이 경우는 도 9(c)의 제1 양자화 세트 Set 0의 복원 레벨로 나타낼 수 있으며, 여기서 세트 0은 양자화 단계 크기(QP)의 모든 짝수 정수배를 포함한다. 신경망 매개변수(13)이 제 2 복원 레벨 세트($setId[k] == 1$)를 사용하여 코딩된 경우 다음 세 가지 경우가 있다: (a) $level[k]$ 가 0과 같으면 n 도 0과 같다; (b) $level[k]$ 가 0보다 크면, n 은 양자화 인덱스 $level[k]$ 의 2배에서 1을 뺀 것과 같다; 및 (c) $level[k]$ 가 0보다 작은 경우, n 은 양자화 인덱스 $level[k]$ 의 2배에 1을 더한 값과 같다. 이것은 부호 함수를 사용하여 지정할 수 있다:

$$\text{sign}(x) = \begin{cases} 1 & : x > 0 \\ 0 & : x = 0 \\ -1 & : x < 0 \end{cases}$$

[0201]

그런 다음 제 2 양자화 세트를 사용하면, 변수 n 은 양자화 인덱스 $level[k]$ 의 2배에서 양자화 인덱스의 부호 함수 $\text{sign}(level[k])$ 를 뺀 것과 같다. 이 경우는 도 9(c)의 제2 양자화 세트 세트 1의 복원 레벨로 나타낼 수 있으며, 여기서 세트 1은 양자화 단계 크기(QP)의 모든 홀수 정수배를 포함한다.

[0202]

변수 n (양자화 단계 크기의 정수 인자 지정)이 결정되면, 복원된 신경망 매개변수 t_k^* 는 n 에 양자화 단계 크기 Δ_k 를 곱하여 얻는다.

[0203]

다시 말해서, 복수의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 2일 수 있고, 본 발명의 실시 예에 따른, 예를 들어 신경망 매개변수(13)를 디코딩 및/또는 인코딩하기 위한 장치는,

[0204]

각각의 신경망 매개변수에 대해 선택된 복원 레벨 세트가 제 1 세트인 경우, 각각의 신경망 매개변수에 대한 양자화 인덱스를 2로 곱하여 각각의 신경망 매개변수에 대한 중간 값을 획득하고;

[0205]

각각의 신경망 매개변수에 대한 선택된 복원 레벨 세트가 제 2 세트이고 각각의 신경망 매개변수에 대한 양자화 인덱스가 0인 경우, 각각의 샘플에 대한 중간 값을 0으로 설정하고;

[0206]

각각의 신경망 매개변수에 대한 선택된 복원 레벨 세트가 제2 세트이고 각각의 신경망 매개변수에 대한 양자화 인덱스가 0보다 큰 경우, 각각의 신경망 매개변수에 대한 양자화 인덱스에 2를 곱하고 곱한 결과에서 1을 빼서 각각의 신경망 매개변수에 대한 중간 값을 획득하고;

[0207]

현재 신경망 매개변수에 대해 선택된 복원 레벨 세트가 제 2 세트이고 각각의 신경망 매개변수에 대한 양자화 인덱스가 0보다 작은 경우, 각각의 신경망 매개변수에 대한 양자화 인덱스에 2를 곱하고 곱한 결과에 1을 더하여 각 신경망 매개변수에 대한 중간 값을 획득함으로써,

[0208]

각각의 신경망 매개변수에 대한 중간 값을 유도하도록 구성될 수 있다.

[0209]

4.3.3 신경망 매개변수의 종속적 복원

[0210]

단락 4.3.1 및 단락 4.3.2에서 상술된 복원 레벨 세트의 선택 외에, 신경망 매개변수 코딩에서 종속 스칼라 양자화의 또 다른 중요한 설계 측면은 정의된 양자화 세트(재건 레벨 세트) 사이를 전환하는 데 사용되는 알고리즘이다. 사용된 알고리즘은 신경망 매개변수(13)의 N 차원 공간에서(따라서 복원된 샘플의 N 차원 공간에서도) 달성할 수 있는 "패킹 밀도"를 결정한다. 더 높은 패킹 밀도는 결국 증가된 코딩 효율성을 초래한다.

[0211]

다음 신경망 매개변수에 대한 복원 레벨 세트를 결정하는 바람직한 방법은 도 11에 도시되어 있는 바와 같이, 양자화 세트의 분할을 기반으로 한다. 도 11은 본 발명의 실시예에 따라 복원 레벨 세트를 2개의 하위 세트로 분할하는 예를 도시한다. 도시된 두 개의 양자화 세트는 도 9(c)의 바람직한 예의 양자화 세트이다. 양자화 세트 0의 두 하위 세트는 "A" 및 "B"를 사용하여 레이블링되고, 양자화 세트 1의 두 하위 세트는 "C" 및 "D"를 사용하여 레이블링된다. 도 11의 양자화 세트는 도 9(c)의 양자화 세트와 동일하다. 두 개(또는 그 이상)의 양자화 세트 각각은 두 개의 하위 세트로 분할된다. 도 11의 바람직한 예의 경우, 제 1 양자화 세트(세트 0으로 표시됨)는 두 개의 하위 세트(A 및 B로 표시됨)으로 분할되고, 제 2 양자화 세트(세트 1로 표시됨)도 2개의 하위 세트(C 및 D로 표시됨)로 분할된다. 유일한 가능성은 아니지만, 각 양자화 세트에 대한 분할은 직접 이웃하는 복원 레벨(및 이에 따라 이웃하는 양자화 인덱스)이 상이한 하위 세트와 연관되는 방식으로 바람직하게 수행된다. 바람직한 실시예에서, 각각의 양자화 세트는 2개의 하위 세트로 분할된다. 도 9에서 양자화 세트를 하위 세트로 분할하게 되면 속이 빈 원과 채워진 원으로 표시된다.

[0212]

도 11 및 도 9(c)에 도시된 특히 바람직한 실시예의 경우, 다음 분할 규칙이 적용된다:

[0213]

- [0214] * 부분 세트 A는 양자화 세트 0의 모든 짝수 양자화 인덱스로 구성된다.
- [0215] * 하위 세트 B는 양자화 세트 0의 모든 홀수 양자화 인덱스로 구성된다.
- [0216] * 부분 세트 C는 양자화 세트 1의 모든 짝수 양자화 인덱스로 구성된다.
- [0217] * 부분 세트 D는 양자화 세트 1의 모든 홀수 양자화 인덱스로 구성된다.

[0218] 사용된 하위 세트는 일반적으로 비트스트림(14) 내부에 명시적으로 표시되지 않는다는 점에 유의해야 한다. 대신, 사용된 양자화 세트(예: 세트 0 또는 세트 1) 및 실제로 전송된 양자화 인덱스(56)를 기반으로 유도될 수 있다. 도 11에 도시된 선호되는 분할의 경우, 하위 세트는 전송된 양자화 인덱스 레벨 및 1의 비트 단위 "and" 연산에 의해 유도될 수 있다. 하위 세트 A는 (level&1)이 0인 세트 0의 모든 양자화 인덱스로 구성되며, 하위 세트 B는 (level&1)이 1인 세트 0의 모든 양자화 인덱스로 구성되며, 하위 세트 C는 (레벨&1)이 0인 세트 1의 모든 양자화 인덱스로 구성되며, 하위 세트 D는 (레벨&1)이 1인 세트 1의 모든 양자화 인덱스로 구성된다.

[0219] 바람직한 실시예에서, 현재 신경망 매개변수(13')를 복원하기 위해 사용되는 양자화 세트(허용 가능한 복원 레벨 세트)는 마지막 2개 이상의 양자화 인덱스(56)와 연관된 하위 세트에 기초하여 결정된다. 두 개의 마지막 부분 세트(마지막 두 개의 양자화 인덱스에 의해 제공됨)이 사용되는 예가 표 1에 나와 있다. 이 표에 의해 지정된 양자화 세트의 결정은 바람직한 실시예를 나타낸다. 다른 실시예에서, 현재 신경망 매개변수(13')에 대한 양자화 세트는 마지막 3개 이상의 양자화 인덱스(56)와 연관된 하위 세트에 의해 결정된다. 계층의 제 1 신경망 매개변수(또는 신경망 매개변수의 하위 세트)에 대해, 이전 신경망 매개변수의 하위 세트에 대한 데이터가 없다(이전 신경망 매개변수가 없기 때문에). 바람직한 실시예에서, 미리 정의된 값이 이러한 경우에 사용된다. 특히 바람직한 실시예에서, 우리는 이용 불가능한 모든 신경망 매개변수에 대한 하위 세트 A를 추론한다. 즉, 제 1 신경망 매개변수를 복원하면, 앞의 두 하위 세트는 "AA"로 추론되므로(또는 3개의 선행 신경망 매개변수가 고려되는 경우 "AAA"), 표 1에 따르면 양자화 세트 0이 사용된다. 제 2 신경망 매개변수의 경우, 바로 앞의 양자화 인덱스의 하위 세트는 그 값에 의해 결정되지만(세트 0은 제 1 신경망 매개변수에 사용되므로 하위 세트는 A 또는 B임), (존재하지 않는) 제 2 마지막 양자화 인덱스에 대한 하위 세트는 A와 동일한 것으로 추론된다. 물론, 존재하지 않는 양자화 인덱스에 대한 기본값을 추론하기 위해 다른 규칙을 사용할 수 있다. 존재하지 않는 양자화 인덱스에 대한 디폴트 하위 세트를 유도하기 위해 다른 구문 요소를 사용하는 것도 가능하다. 다른 대안으로서, 초기화를 위해 신경망 매개변수(13)의 이전 세트의 마지막 양자화 인덱스(56)를 사용하는 것도 가능하다.

[0220] 표 1: 본 발명의 실시예에 따른 2개의 마지막 양자화 인덱스와 연관된 하위 세트에 기초한 다음 신경망 매개변수에 사용되는 양자화 세트(사용 가능한 복원 레벨 세트) 결정의 예

[0221] 하위 세트는 왼쪽 표 열에 표시되고; 그들은 사용된 양자화 세트(두 개의 마지막 양자화 인덱스에 대해)와 (양자화 인덱스의 패리티에 의해 결정될 수 있는) 소위 경로에 의해 고유하게 결정된다. 양자화 세트와 괄호 안에 있는 하위 세트의 경로는 왼쪽에서 제 2 열에 나열된다. 제 3 열은 연관된 양자화 세트를 지정한다. 마지막 열에는, 양자화 세트를 결정하는 프로세스를 단순화하는 데 사용할 수 있는 소위 상태 변수의 값이 표시된다.

표 1

두 마지막 양자화 인덱스의 하위 세트	두 마지막 양자화 인덱스에 대한 양자화 세트 및 경로(괄호안에 주어져)	현재 신경망 매개변수에 대한 양자화 세트	상태 변수
A A	0(0), 0(0)	0	0
A B	0(0), 0(1)	0	0
A C	0(0), 1(0)	1	1
A D	0(0), 1(1)	1	1
B A	0(1), 0(0)	1	1
B B	0(1), 0(1)	1	1
B C	0(1), 1(0)	0	0
B D	0(1), 1(1)	0	0
C A	1(0), 0(0)	0	2
C B	1(0), 0(1)	0	2
C C	1(0), 1(0)	1	3
C D	1(0), 1(1)	1	3
D A	1(1), 0(0)	1	3
D B	1(1), 0(1)	1	3

D C	1(1), 1(0)	0	2
D D	1(1), 1(1)	0	2

- [0223] 양자화 인덱스(56)의 하위 세트(A, B, C 또는 D)는 사용된 양자화 세트(세트 0 또는 세트 1) 및 양자화 세트 내의 사용된 하위 세트(예를 들어, 세트 0의 경우 A 또는 B, 세트 1의 경우 C 또는 D)에 의해 결정된다는 점에 유의한다. 양자화 세트 내에서 선택된 부분 세트를 경로라고도 한다(아래에서 설명하는 바와 같이 종속 양자화 프로세스를 격자 구조로 표현하면 경로를 지정하기 때문). 우리 규칙에서 경로는 0 또는 1과 같다. 그런 다음 부분 세트 A는 세트 0의 경로 0에 해당하고 부분 세트 B는 세트 0의 경로 1에 해당하고 부분 세트 C는 세트 1의 경로 0에 해당하고 부분 세트 D는 세트 1의 경로 1에 해당한다. 따라서 다음 신경망 매개변수에 대한 양자화 세트도 또한 2개(또는 그 이상)의 마지막 양자화 인덱스와 연관되는 양자화 세트(세트 0 또는 세트 1)와 경로(경로 0 또는 경로 1)에 의해 고유하게 결정된다. 표 1에서 관련 양자화 세트 및 경로는 제 2 열에 지정된다.
- [0224] 경로는 이진 함수와 같은 간단한 산술 연산으로 종종 결정될 수 있다는 점에 유의해야 한다. 예를 들어, 도 11에 표시된 구성의 경우, 경로는 다음과 같이 지정된다:
- [0225] $path = (level[k] \& 1)$
- [0226] 여기에서 $level[k]$ 는 양자화 인덱스(가중치 레벨)(56)을 나타내고 연산자 $\&$ 는 비트 단위 "and"를 지정한다(2-보수 정수 산술에서).
- [0227] 다시 말해서, 복수(50)의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 세트 0 및 세트 1, 2일 수 있으며, 예를 들어 본 발명의 실시예에 따라 신경망 매개변수(13)를 디코딩하기 위한 장치는 각각의 신경망 매개변수에 대한 복원 레벨의 선택된 세트 및 각각의 신경망 매개변수에 대한 양자화 인덱스의 이진 함수에 기초하여 각각의 신경망 매개변수에 대한 하위 세트 인덱스를 유도하여, 결과적으로 네 가지 가능한 값, 예를 들어, 하위 세트 인덱스의 경우 A, B, C 또는 D가 생기게 하고, 현재 신경망 매개변수(13')에 대해, 이전에 디코딩된 신경망 매개변수에 대한 하위 세트 인덱스에 따라, 복수(50)의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 선택(54)하도록 구성된다.
- [0228] 본 발명에 따른 추가 실시예는 예를 들어 표 1의 제 1 열과 같이, 바로 이전에 디코딩된 다수의 신경망 매개변수에 대한 하위 세트 인덱스에 의존하는 선택 규칙을 사용하여, 현재 신경망 매개변수(13')에 대해, 복수의 복원 레벨 세트 5) 중에서 복원 레벨 세트(48)를 선택(54)하고, 신경망 매개변수의 전체 또는 일부에 대해 선택 규칙을 사용하도록 구성된 장치를 포함한다.
- [0229] 추가 구현예에 따르면, 선택 규칙이 의존하는 직전에 디코딩된 신경망 매개변수의 수는 2이고, 예를 들어 표 1에 나타난 바와 같이, 2개의 마지막 양자화 인덱스의 하위 세트가다.
- [0230] 추가적인 실시예에 따르면, 각 신경망 매개변수에 대한 하위 세트 인덱스는 각 신경망 매개변수에 대해 선택된 복원 레벨 세트 및 각각의 신경망 매개변수에 대한, 예를 들어 $path = (level[k] \& 1)$ 사용하는 양자화 인덱스의 패리티를 기반으로 유도된다.
- [0231] 각각, 실시예에 따른 신경망 매개변수 인코딩 장치(13)에 대해, 복수의 복원 레벨 세트(52)의 복원 레벨 세트(52)의 수는 예를 들어 세트 0 및 세트 1로, 2일 수 있고, 장치는 각각의 신경망 매개변수에 대한 복원 레벨의 선택된 세트 및 각각의 신경망 매개변수에 대한 양자화 인덱스의 이진 함수에 기초하여, 각각의 신경망 매개변수에 대한 하위 세트 인덱스를 유도하여, 결과적으로 하위 세트 인덱스에 대해 4개의 가능한 값, 예를 들어, A, B, C 및 D가 결과되고, 이전에 인코딩된 신경망 매개변수에 대한 하위 세트 인덱스에 따라, 현재 신경망 매개변수(13')에 대해, 복수(50)의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 선택(54)하도록 구성된다.
- [0232] 본 발명에 따른 추가 실시예는 예를 들어 표 1의 제 1 열과 같이, 바로 이전에 인코딩된 많은 신경망 매개변수에 대한 하위 세트 인덱스에 의존하는 선택 규칙을 사용하여, 현재 신경망 매개변수(13')에 대해, 복수(50)의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 선택(54)하고, 신경망 매개변수의 전체 또는 일부에 대해 선택 규칙을 사용하도록 구성된 장치를 포함한다.
- [0233] 추가 구현예에 따르면, 선택 규칙이 의존하는 바로 이전에 인코딩된 신경망 매개변수의 수는 2이고, 예를 들어 표 1에 나타난 바와 같이, 2개의 마지막 양자화 인덱스의 하위 세트이다.
- [0234] 추가적인 실시예에 따르면, 각각의 신경망 매개변수에 대한 하위 세트 인덱스는 예를 들어, 각각의 신경망 매개

변수에 대한 복원 레벨의 선택된 세트 및 각각의 신경망 매개변수에 대한 예를 들어 양자화 인덱스의 경로=(level[k] & 1)를 사용하는 양자화 인덱스의 패리티에 기초하여 유도된다.

[0235] 양자화 세트(52)(세트 0 및 세트 1) 사이의 전이는 또한 상태 변수에 의해 우아하게 표현될 수 있다. 이러한 상태 변수의 예는 표 1의 마지막 열에 나와 있다. 이 예에서 상태 변수에는 4개의 가능한 값(0, 1, 2, 3)이 있다. 한편, 상태 변수는 현재 신경망 매개변수(13')에 사용되는 양자화 세트를 지정한다. 표 1의 바람직한 예에서, 양자화 세트 0은 상태 변수가 0 또는 2인 경우에만 사용되며, 양자화 세트 1은 상태 변수가 1 또는 3인 경우에만 사용된다. 반면에 상태 변수는 양자화 세트 간의 가능한 전환도 지정한다. 상태 변수를 사용하면 표 1의 규칙을 더 작은 상태 전이 표로 설명할 수 있다. 예를 들어, 표 2는 표 1에 주어진 규칙에 대한 상태 전이 표를 지정한다. 이것은 바람직한 실시예를 나타낸다. 현재 상태가 주어지면, 현재 신경망 매개변수(제 2 열)에 대한 양자화 세트를 지정했다. 선택된 양자화 인덱스(56)와 관련된 경로를 기반으로 상태 전환을 추가로 지정한다(경로는 양자화 세트가 주어진 경우 사용된 하위 세트 A, B, C 또는 D를 지정함). 상태 변수의 개념을 사용하게 되면, 실제로 선택된 하위 세트를 추적할 필요는 없다는 것에 유의한다. 계층에 대한 신경망 매개변수를 복원할 때, 상태 변수를 업데이트하고 사용된 양자화 인덱스의 경로를 결정하는 것으로 충분하다.

[0236] 표 2: 본 발명의 실시예에 따른, 4개의 상태를 갖는 구성에 대한 상태 전이 표의 바람직한 예

표 2

현재 상태	현재 계수에 대한 양자화 세트	다음 상태	
		경로 0	경로 1
0	0	0	1
1	1	2	3
2	0	1	0
3	1	3	2

[0238] 즉, 장치, 예를 들어 실시예에 따른 신경망 매개변수를 디코딩하기 위한 장치는 현재 신경망 매개변수(13')에 대해 현재 신경망 매개변수(13')와 연관된 상태에 따라 복수(50)의 복원 레벨 세트(52) 중 양자화 레벨 세트(48)를 결정하고, 직전 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상태를 업데이트함으로써, 현재 신경망 매개변수(13')에 대해, 상태 전이 프로세스에 의해 복수(50)의 복원 레벨 세트(52) 중에서 양자화 레벨 세트(48)를 선택하도록(54) 구성될 수 있다.

[0239] 각각, 실시예에 따른 신경망 매개변수 인코딩 장치(13)에 대해, 상기 장치는 현재 신경망 매개변수(13')에 대해 현재 신경망 매개변수(13')와 연관된 상태에 따라 복수의 복원 레벨 세트(52) 중 복원 레벨 세트(48)를 결정함으로써, 및 직전 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스(58)에 따라 후속 신경망 매개변수에 대한 상태를 업데이트함으로써, 상태 전환 프로세스를 통해 현재 신경망 매개변수(13')에 대해 복수(50)의 복원 레벨 세트(52) 중에서 복원 레벨 세트(48)를 선택(54)하도록 구성될 수 있다.

[0240] 본 발명의 바람직한 실시예에서, 경로는 양자화 인덱스의 패리티에 의해 지정된다. level[k]가 현재 양자화 인덱스인 경우, 이것은 다음에 따라 결정할 수 있다:

[0241]
$$\text{path} = (\text{level}[k] \& 1)$$

[0242] 여기에서 연산자 &는 2-보수 정수 산술에서 비트 단위 "and"를 나타낸다.

[0243] 즉, 예를 들어 실시예에 따른 신경망 매개변수를 디코딩하기 위한 장치는 직전 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(58)의 이전 함수를 사용하는 후속 신경망 매개변수에 대해, 예를 들어, 표 2에 따라 상태를 업데이트하도록 구성될 수 있다.

[0244] 또한, 실시예에 따른 장치는 예를 들어, 바로 앞의 신경망 매개변수에 대해 데이터 스트림(14)으로부터 디코딩되는, path=(level[k] & 1)를 사용하여, 양자화 인덱스(58)의 패리티를 사용하여 후속 신경망 매개변수에 대한 상태를 업데이트하도록 구성될 수 있다.

[0245] 각각, 실시예에 따른 신경망 매개변수 인코딩 장치(13)에 대해, 상기 장치는 직전의 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스(58)의 이전 함수를 사용하여 후속 신경망 매개변수에 대한 상태를 업데이트하도록 구성될 수 있다.

[0246] 또한, 예를 들어 실시예에 따라 신경망 매개변수(13)를 인코딩하기 위한 장치는, 직전의 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스(58)의 패리티를 사용하는 후속 신경망 매개변수에 대해, 예를 들어, 표 2에 따라 상태를 업데이트하도록 구성될 수 있다.

[0247] 바람직한 실시예에서, 4개의 가능한 값을 갖는 상태 변수가 사용된다. 다른 실시예에서, 상이한 수의 가능한 값을 갖는 상태 변수가 사용된다. 상태 변수에 대해 가능한 값의 수가 2의 정수 거듭제곱, 즉 4, 8, 16, 32, 64 등을 나타내는 상태 변수가 특히 중요하다. 바람직한 구성(표 1 및 표 2 참조)에서, 4개의 가능한 값을 갖는 상태 변수는 현재 양자화 세트가 2개의 마지막 양자화 인덱스의 하위 세트에 의해 결정되는 접근 방식과 동일하다는 것에 유의한다. 8개의 가능한 값을 갖는 상태 변수는 현재 양자화 세트가 3개의 마지막 양자화 인덱스의 하위 세트에 의해 결정되는 유사한 접근 방식에 해당한다. 16개의 가능한 값을 가진 상태 변수는 현재 양자화 세트가 마지막 4개 양자화 인덱스 등의 하위 세트에 의해 결정되는 접근 방식에 해당한다. 2의 정수 거듭제곱과 같은 가능한 값의 수를 가진 상태 변수를 사용하는 것이 일반적으로 바람직하지만, 실시예는 이 설정으로 제한되지 않는다.

[0248] 특히 바람직한 구현예에서, 8개의 가능한 값(0, 1, 2, 3, 4, 5, 6, 7)이 있는 상태 변수가 사용된다. 바람직한 실시예 표 3에서, 양자화 세트 0은 상태 변수가 0, 2, 4 또는 6인 경우에만 사용되고, 양자화 세트 1은 상태 변수가 1, 3, 5 또는 7인 경우에만 사용된다.

[0249] 표 3: 실시예에 따른, 8개의 상태를 갖는 구성에 대한 상태 전이 표의 바람직한 예.

표 3

현재 상태	현재 계수에 대한 양자화 계수	다음 상태	
		경로 0	경로 1
0	0	0	2
1	1	7	5
2	0	1	3
3	1	6	4
4	0	2	0
5	1	5	7
6	0	3	1
7	1	4	6

[0251] 즉, 본 발명의 실시예에 따르면, 상태 전환 프로세스는 4개 또는 8개의 가능한 상태 사이를 전환하도록 구성된다.

[0252] 또한, 실시예에 따른 신경망 매개변수(13)를 디코딩/인코딩하기 위한 장치는, 상태 전이 프로세스에서, 짝수개의 가능한 상태 사이에서 전이하도록 구성되고 복수(50)의 복원 레벨 세트(52) 중 복원 레벨 세트(52)의 수가 2이고, 여기에서 현재 신경망 매개변수(13')에 대해 현재 신경망 매개변수(13')와 연관된 상태에 따라 양자화 세트(52) 중 양자화 레벨의 세트(48)를 결정하는 단계는, 상태가 짝수개의 가능한 상태의 전반부에 속하는 경우 복수(50)의 복원 레벨 세트(52) 중에서 설정된 제1 복원 레벨, 및 상태가 짝수개의 가능한 상태의 후반부에 속하는 경우 복수(50)의 복원 레벨 세트(52) 중에서 설정된 제2 복원 레벨을 결정한다.

[0253] 예를 들어, 추가 실시예에 따라 신경망 매개변수(13)를 디코딩하기 위한 장치는 바로 선행하는 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(58)의 패리티 및 상태의 조합을 후속 신경망 매개변수와 연관된 추가 상태로 매핑하는 전환 표를 통해, 상태 업데이트를 수행하도록 구성될 수 있다.

[0254] 각각, 실시예에 따른 신경망 매개변수 인코딩 장치(13)는 바로 선행하는 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스(58)의 패리티와 상태의 조합을 후속 신경망 매개변수와 연관된 추가 상태에 매핑하는 전환 표를 통해 상태 업데이트를 수행하도록 구성될 수 있다.

[0255] 상태 전이의 개념을 사용하여, 현재 상태 및 따라서 현재 양자화 세트는 이전 상태(복구 순서로) 및 이전 양자화 인덱스(56)에 의해 고유하게 결정된다. 그러나 유한 세트(예: 계층)의 제 1 신경망 매개변수(13)의 경우에, 이전 상태 및 이전 양자화 인덱스가 없다. 따라서 계층의 제 1 신경망 매개변수에 대한 상태가 고유하게 정의되어야 한다. 다양한 가능성이 있다. 선호하는 선택은 다음과 같다.

[0256] * 계층의 제 1 상태는 항상 미리 정의된 고정 값과 동일하게 설정된다. 바람직한 실시예에서, 제1 상태는 0과

동일하게 설정된다.

- [0257] * 제 1 상태의 값은 비트스트림(14)의 일부로 명시적으로 전송된다. 여기에는 가능한 상태 값의 하위 세트만 해당 구문 요소로 표시될 수 있는 접근 방식이 포함된다.
- [0258] * 제 1 상태의 값은 계층에 대한 다른 구문 요소를 기반으로 유도된다. 이것은 대응하는 구문 요소(또는 구문 요소)가 디코더에 다른 측면을 시그널링하기 위해 사용되더라도, 종속 스칼라 양자화에 대해 제 1 상태를 유도하는 데 추가로 사용되는 것을 의미한다.
- [0259] 종속 스칼라 양자화에 대한 상태 전이의 개념은 디코더에서 신경망 매개변수(13)의 복원을 위한 낮은 복잡성 구현을 허용한다. 단일 계층의 신경망 매개변수의 복원 프로세스에 대한 바람직한 예는 C 스타일 의사 코드를 사용하여 도 12에 나와 있다. 도 12는 본 발명의 실시예에 따른 계층에 대한 신경망 매개변수(13)의 복원 프로세스에 대한 바람직한 예를 나타내는 의사 코드의 예를 도시한다. 대안적으로, 양자화 단계 크기를 사용하거나 대안적으로 코드북을 사용하여 양자화 인덱스의 유도 및 복원된 값의 유도는 별도의 루프에서 차례로 수행될 수 있다. 즉, 다시 말해, "n"의 유도 및 상태 업데이트는 제 1 루프에서 수행되고, "trec"의 유도는 다른 별도의 제 2 루프에서 수행된다. 어레이 레벨(210)은 계층에 대해 전송된 신경망 매개변수 레벨(양자화 인덱스(56))을 나타내고 어레이 trec(220)은 대응하는 복원된 신경망 매개변수(13)를 나타낸다. 현재 신경망 매개변수(13')에 적용되는 양자화 단계 크기 $\Delta_k(QP)$ 는 quant_step_size[k]로 표시된다. 2d 표 sttab(230)은 표 1, 2 및/또는 3 중 어느 하나에 따라, 상태 전이 표를 지정하고, 표 setId(240)는 상태(250)와 연관된 양자화 세트를 지정한다.
- [0260] 도 12의 의사 코드에서 인덱스 k는 신경망 매개변수의 복원 순서를 지정한다. 마지막 인덱스 layerSize는 마지막으로 복원된 신경망 매개변수의 복원 인덱스를 지정한다. 변수 layerSize는 계층의 신경망 매개변수의 수와 동일하게 설정될 수 있다. 각각의 단일 신경망 매개변수에 대한 복원 프로세스는 도 10의 예와 동일한다. 도 10의 예와 같이, 양자화 인덱스는 level[k](210)으로 표시되고 연관된 복원된 신경망 매개변수는 trec[k](220)로 표시된다. 상태 변수는 상태(210)으로 표시된다. 도 12의 예에서 상태는 계층 시작 시 0으로 설정된다. 그러나 위에서 논의한 바와 같이, 다른 초기화(예: 일부 구문 요소의 값을 기반으로 함)가 가능하다. 1d 표 setId[] (240)는 상태 변수의 다른 값과 연관된 양자화 세트를 지정하고 2d 표 sttab[][] (230)은 주어진 현재 상태(제 1 인수) 및 경로(제 2 인수)가 주어진 상태 전이를 지정한다. 예제에서, 경로는 양자화 인덱스의 패리티(비트 단위 및 연산자 &를 사용)에 의해 제공되지만 다른 개념도 가능하다. 표에 대한 C 스타일 구문의 예는 도 13 및 도 14에 나와 있다(이러한 표는 표 2 및 표 3과 동일하다. 즉, 표 2 및 표 3의 표현을 제공할 수 있다).
- [0261] 도 13은 본 발명의 실시예에 따른 상태(250)와 연관된 양자화 세트를 지정하는, 상태 전이 표 sttab(230) 및 표 setId(240)에 대한 바람직한 예를 도시한다. C 스타일 구문으로 제공된 표은 표 2에 지정된 표를 나타낸다.
- [0262] 도 14는 본 발명의 실시예에 따라 상태(250)와 연관된 양자화 세트를 지정하는, 상태 전이 표 sttab(230) 및 표 setId(240)에 대한 바람직한 예를 도시한다. C 스타일 구문으로 제공된 표은 표 3에 지정된 표을 나타낸다.
- [0263] 다른 실시예에서, 0과 동일한 모든 양자화 인덱스(56)는 상태 전이 및 종속 복원 프로세스로부터 제외된다. 양자화 인덱스(56)가 0인지 아닌지에 대한 정보는 단지 신경망 매개변수(13)를 0 및 0이 아닌 신경망 매개변수로 분할하기 위해 사용된다. 종속 스칼라 양자화를 위한 복원 프로세스는 0이 아닌 양자화 인덱스(56)의 정렬된 세트에만 적용된다. 0과 같은 양자화 인덱스와 관련된 모든 신경망 매개변수는 단순히 0으로 설정된다. 해당 의사 코드는 도 15에 나와 있다. 도 15는 본 발명의 실시예에 따라, 0과 동일한 양자화 인덱스는 상태 전이 및 종속 스칼라 양자화에서 제외되는, 신경망 매개변수 레벨에 대한 대안적인 복원 프로세스를 예시하는 의사 코드를 도시한다.
- [0264] 종속 양자화의 상태 전이는 도 16과 같이 격자 구조를 사용하여 나타낼 수도 있다. 도 16은 본 발명의 실시예에 따른 격자 구조로서의 종속 스칼라 양자화에서의 상태 전이의 예를 도시한다. 가로축은 복원 순서에 따라 다른 신경망 매개변수(13)를 나타낸다. 수직축은 종속 양자화 및 복원 프로세스에서 다양한 가능한 상태(250)를 나타낸다. 표시된 연결은 다양한 신경망 매개변수의 상태 간에 사용 가능한 경로를 지정한다. 이 도면에 표시된 격자는 표 2에 지정된 상태 전환에 해당한다. 각 상태(250)에 대해, 현재 신경망 매개변수(13')에 대한 상태를 복원 순서로 다음 신경망 매개변수(13)에 대한 두 가지 가능한 상태와 연결하는 두 개의 경로가 있다. 경로는 경로 0과 경로 1로 레이블이 지정되며, 이 숫자는 위에서 소개한 경로 변수에 해당한다(바람직한 실시예에서, 그 경로 변수는 양자화 인덱스의 패리티와 동일하다). 각 경로에는 양자화 인덱스에 대한 하위 세트(A, B, C 또는 D)가 고유하게 지정된다는 것에 유의한다. 도 16에서, 하위 세트는 괄호 안에 지정되어 있다. 초기 상태(예: 상태 0)가 주어지면, 격자를 통한 경로는 전송된 양자화 인덱스(56)에 의해 고유하게 지정된다.

- [0265] 도 16의 예에서, 상태(0, 1, 2, 3)는 다음 속성을 갖는다.
- [0266] * 상태 0: 이전 양자화 인덱스 레벨[k-1]은 세트 0의 복원 레벨을 지정하고 현재 양자화 인덱스 레벨[k]은 세트 0의 복원 레벨을 지정한다.
- [0267] * 상태 1: 이전 양자화 인덱스 레벨[k-1]은 세트 0의 복원 레벨을 지정하고 현재 정량 인덱스 레벨[k]는 세트 1의 복원 레벨을 지정한다.
- [0268] * 상태 2: 이전 양자화 인덱스 레벨[k-1]은 세트 1의 복원 레벨을 지정하고 현재 정량화 인덱스 레벨[k]은 세트 0의 복원 레벨을 지정한다.
- [0269] * 상태 3: 이전 양자화 인덱스 레벨[k-1]은 세트 1의 복원 레벨을 지정하고 현재 정량 인덱스 레벨[k]는 세트 1의 복원 레벨을 지정한다.
- [0270] 격자는 소위 기본 격자 셀의 연결로 구성된다. 이러한 기본 격자 셀의 예가 도 17에 나와 있다. 도 17은 본 발명의 실시예에 따른 기본 격자 셀의 예를 도시한다. 본 발명은 4개의 상태(250)를 갖는 격자로 제한되지 않는다는 점에 유의한다. 다른 실시예에서, 격자는 더 많은 상태(250)를 가질 수 있다. 특히, 2의 정수 거듭제곱을 나타내는 임의의 수의 상태가 적합하다. 특히 바람직한 실시예에서 상태(250)의 수는 8과 동일하다. 표 3과 유사하다. 격자가 2개 이상의 상태(250)를 가지고 있더라도, 현재 신경망 매개변수(13')에 대한 각 노드는 일반적으로 이전 신경망 매개변수(13)에 대한 두 가지 상태와 다음 신경망 매개변수(13)의 두 가지 상태와 연결된다. 그러나 노드가 이전 신경망 매개변수의 두 개 이상의 상태 또는 다음 신경망 매개변수의 두 개 이상의 상태와 연결될 수도 있다. 완전히 연결된 격자(각 상태(250)는 이전 신경망 매개변수(13)의 모든 상태(250) 및 다음 신경망 매개변수(13)의 모든 상태(250)와 연결됨)는 독립 스칼라 양자화에 해당한다는 것에 유의한다.
- [0271] 바람직한 실시예에서, 초기 상태는 자유롭게 선택될 수 없다(이 결정을 디코더에 전송하려면 약간의 부가 정보 속도가 필요하기 때문에). 대신에, 초기 상태가 미리 정의된 값으로 설정되거나 그 값이 다른 구문 요소를 기반으로 유도된다. 이 경우, 모든 경로 및 상태(250)가 제1 신경망 매개변수에 이용 가능한 것은 아니다. 4 상태 trellis의 예로서, 도 18은 초기 state가 0인 경우에 대한 trellis 구조를 도시한다. 도 18은 본 발명의 실시예에 따른 8개의 신경망 매개변수의 종속 스칼라 양자화를 위한 격자의 예를 도시한다. 제 1 상태(왼쪽)는 이 예에서 0으로 설정되는, 초기 상태를 나타낸다.
- [0272] 4.4 엔트로피 코딩
- [0273] 종속 양자화에 의해 획득된 양자화 인덱스는 엔트로피 코딩 방법을 사용하여 인코딩된다. 이를 위해 모든 엔트로피 코딩 방법을 적용할 수 있다. 본 발명의 바람직한 실시예에서, 컨텍스트 적응 이진 산술 코딩(CABAC)를 사용하여 섹션 2.2에 따른 엔트로피 코딩 방법이 적용된다(엔코더 방법은 섹션 2.2.1, 디코더 방법은 섹션 2.2.2 참조). 이를 위해, 예를 들어 도 5(이진화)에 도시된 바와 같이, 양자화 인덱스를 절대 값으로 전송하기 위해 비 이진은 먼저 일련의 이진 결정(소위 빈)에 매핑된다.
- [0274] 본 명세서에서 설명된 모든 개념은 단락 3에서 방법 및 관련 개념(특히 컨텍스트 모델링과 관련하여)과 결합될 수 있다는 것에 유의한다는 점에 유의한다.
- [0275] 4.4.1 종속 스칼라 양자화를 위한 컨텍스트 모델링
- [0276] 종속 스칼라 양자화의 주요 측면은 신경망 매개변수(13)에 대해 다른 세트의 허용 가능한 복원 레벨(양자화 세트라고도 함)이 있다는 것이다. 현재 신경망 매개변수(13')에 대한 양자화 세트는 이전 신경망 매개변수에 대한 양자화 인덱스(56)의 값에 기초하여 결정된다. 도 11의 바람직한 예를 고려하고 두 개의 양자화 세트를 비교하면, 0과 동일한 복원 레벨과 이웃하는 복원 레벨 사이의 거리는 세트 1에서보다 세트 0에서 더 크다는 것이 명백하다. 따라서, 양자화 인덱스(56)가 0일 확률은 세트 0이 사용되면 더 크고 세트 1이 사용되면 더 작다. 바람직한 실시예에서, 이 효과는 현재 양자화 인덱스에 사용되는 양자화 세트(또는 상태)를 기반으로 코드워드 표 또는 확률 모델을 전환함으로써 엔트로피 코딩에서 이용된다.
- [0277] 코드워드 표 또는 확률 모델의 적절한 전환을 위해, 현재 양자화 인덱스 (또는 현재 양자화 인덱스의 대응하는 이진 결정)를 엔트로피 디코딩할 때 모든 선행 양자화 인덱스의 경로(사용된 양자화 세트의 하위 세트과의 연결)를 알아야 한다. 따라서, 신경망 매개변수(13)는 복원 순서로 코딩될 필요가 있다. 따라서, 바람직한 실시예에서, 신경망 매개변수(13)의 코딩 순서는 이들의 복원 순서와 동일하다. 그런 측면 말고도, 섹션 2.2.1에 지정된 것과 같은 양자화 인덱스(56)의 임의의 코딩/복원 순서는, 다른 고유하게 정의된 순서이다.

- [0278] 즉, 본 발명에 따른 실시예는, 예를 들어 이전에 인코딩된 신경망 매개변수의 양자화 인덱스에 추가로 의존하는 확률 모델을 사용하여 신경망 매개변수를 인코딩하기 위한 장치를 포함한다.
- [0279] 각각, 본 발명에 따른 실시예는, 예를 들어 이전에 디코딩된 신경망 매개변수의 양자화 인덱스에 추가로 의존하는 확률 모델을 사용하여 신경망 매개변수를 디코딩하기 위한 장치를 포함한다.
- [0280] 절대 레벨에 대한 빈의 적어도 일부는 일반적으로 적응 확률 모델(컨텍스트라고도 함)을 사용하여 코딩된다. 본 발명의 바람직한 실시예에서, 하나 이상의 빈의 확률 모델은 해당하는 신경망 매개변수에 대해, 양자화 세트(또는 더 일반적으로, 예를 들어 표 1-3 중 어느 하나에 따른 관계를 갖는 해당 상태 변수)를 기반으로 선택된다. 선택된 확률 모델은 이미 전송된 양자화 인덱스(56)의 다중 매개변수 또는 속성에 의존할 수 있지만, 매개변수 중 하나는 코딩되는 양자화 인덱스에 적용되는 양자화 세트 또는 상태이다.
- [0281] 다시 말해서, 실시예에 따라 예를 들어 신경망 매개변수(13)를 인코딩하기 위한 장치는, 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨의 상태 또는 세트(48)에 따라, 복수의 확률 모델 중에서 확률 모델의 하위 세트를 사전 선택하고, 이전에 인코딩된 신경망 매개변수의 양자화 인덱스에 따라(121) 확률 모델의 하위 세트 중에서 현재 신경망 매개 변수에 대한 확률 모델을 선택하도록 구성된다.
- [0282] 실시예에 따른, 예를 들어 신경망 매개변수(13)를 디코딩하기 위한 각각의 장치는 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨의 상태 또는 세트(48)에 따라, 복수의 확률 모델 중에서 확률 모델의 하위 세트를 사전 선택하고, 이전에 디코딩된 신경망 매개변수의 양자화 인덱스에 따라(121) 확률 모델의 하위 세트에서 현재 신경망 매개 변수에 대한 확률 모델을 선택하도록 구성될 수 있다.
- [0283] 예를 들어, 도 9의 맥락에서 설명된 바와 같은 본 발명의 개념과 조합하여, 예를 들어 본 발명에 따른 신경망 매개변수(13)의 인코딩 및/또는 디코딩을 위한 실시예는, 현재 신경망 매개변수(13')에 대해 선택된 복원 레벨의 상태 또는 세트(48)에 따라, 제1 상태 또는 복원 레벨 세트에 대해 사전 선택된 하위 세트가 임의의 다른 상태 또는 복원 레벨 세트에 대해 사전 선택된 하위 세트와 분리되도록 하는 방식으로, 다수의 확률 모델 중 확률 모델의 하위 세트를 사전 선택하도록 구성된 장치를 포함한다.
- [0284] 특히 바람직한 구현예에서, 계층의 양자화 인덱스를 전송하기 위한 구문은 양자화 인덱스가 0인지 또는 0이 아닌지, 예를 들어 앞서 언급한 `sig_flag`인지를 지정하는 빈을 포함한다. 이 빈을 코딩하는 데 사용되는 확률 모델은 두 개 이상의 확률 모델 세트 중에서 선택된다. 사용된 확률 모델의 선택은 대응하는 양자화 인덱스(56)에 적용하는 양자화 세트(즉, 복원 레벨 세트)에 따라 다르다. 본 발명의 다른 실시예에서, 사용된 확률 모델은 현재 상태 변수에 따라 다르다(상태 변수는 사용된 양자화 세트를 의미한다).
- [0285] 추가 실시예에서, 계층의 양자화 인덱스를 전송하기 위한 구문은 양자화 인덱스가 0보다 크거나 0보다 작은지, 예를 들어 앞서 언급한 `sign_flag`인지를 지정하는 빈을 포함한다. 즉, 빈은 양자화 인덱스의 부호를 나타낸다. 사용된 확률 모델의 선택은 해당 양자화 인덱스에 적용하는 양자화 세트(즉, 복원 레벨 세트)에 따라 다르다. 다른 실시예에서, 사용된 확률 모델은 현재 상태 변수에 의존한다(상태 변수는 사용된 양자화 세트를 의미한다).
- [0286] 추가 실시예에서, 양자화 인덱스를 전송하기 위한 구문은 양자화 인덱스(신경망 매개변수 레벨)의 절대값이 X , 예를 들어 앞서 언급한 `abs_level_greater_X`(자세한 내용은 섹션 0 참조)보다 큰지 여부를 지정하는 빈을 포함한다. 이 빈을 코딩하는 데 사용되는 확률 모델은 두 개 이상의 확률 모델 세트 중에서 선택된다. 사용된 확률 모델의 선택은 대응하는 양자화 인덱스(56)에 적용하는 양자화 세트(즉, 복원 레벨 세트)에 따라 다르다. 다른 실시예에서, 사용된 확률 모델은 현재 상태 변수에 의존한다(상태 변수는 사용된 양자화 세트를 의미한다).
- [0287] 본 명세서에 논의된 실시예의 한 유리한 측면은 신경망 매개변수(13)의 중속 양자화가 엔트로피 코딩과 결합된다는 점이고, 여기에서 (양자화 레벨이라고도 하는) 양자화 인덱스의 이진 표현의 하나 이상의 빈에 대한 확률 모델의 선택은 양자화 세트(허용 가능한 복원 레벨 세트) 또는 현재 양자화 인덱스에 대한 해당 상태 변수에 따라 다르다. 양자화 세트(52)(또는 상태 변수)는 코딩 및 복원 순서에서 이전 신경망 매개변수에 대해, 양자화 인덱스(56)(또는 양자화 인덱스를 나타내는 빈의 하위 세트)에 의해 제공된다.
- [0288] 바람직한 구현예에서, 설명된 확률 모델 선택은 다음 엔트로피 코딩 측면 중 하나 이상과 결합된다:
- [0289] * 양자화 인덱스의 절대값은 적응 확률 모델을 사용하여 코딩된 다수의 빈, 및 적응 코딩된 빈이 아직 절대값을 완전히 지정하지 않은 경우, 산술 코딩 엔진(모든 빈에 대해 `pmf`(예: 확률 질량 함수)(0.5, 0.5)가 있는 비적응 확률 모델)의 바이패스 모드에서 코딩된 접미사 부분으로 구성된 이진화 방식을 사용하여 전송된다. 바람직한

실시예에서, 접미사 부분에 사용되는 이진화는 이미 전송된 양자화 인덱스의 값에 따라 다르다.

- [0290] * 양자화 인덱스의 절대값에 대한 이진화는 양자화 인덱스가 0과 같지 않은지 여부를 지정하는 적응적으로 코딩된 빈을 포함한다. 이 빈을 코딩하는 데 사용되는 확률 모델(컨텍스트라고 함)은 후보 확률 모델 세트 중에서 선택된다. 선택된 후보 확률 모델은 양자화 세트(허용 가능한 복원 레벨 세트) 또는 현재 양자화 인덱스(56)에 대한 상태 변수에 의해 결정될 뿐만 아니라, 또한 계층에 대해 이미 전송된 양자화 인덱스에 의해서도 결정된다. 바람직한 실시예에서, 양자화 세트(또는 상태 변수)는 사용 가능한 확률 모델의 하위 세트(컨텍스트 세트라고도 함)를 결정하고, 이미 코딩된 양자화 인덱스의 값은 이 하위 세트(컨텍스트 세트) 내에서 사용된 확률 모델을 결정한다.
- [0291] 일 실시예에서, 컨텍스트 세트 내에서 사용된 확률 모델은 현재 신경망 매개변수의 로컬 이웃, 예를 들어, 2.2.3에 설명된 템플릿에서, 이미 코딩된 양자화 인덱스의 값을 기반으로 결정된다. 이하, 로컬 이웃의 양자화 인덱스 값을 기반으로 유도될 수 있고 다음에 사전 결정된 컨텍스트 세트의 확률 모델을 선택하는 데 사용할 수 있는 몇 가지 예시적인 측정값이 나열된다:
- [0292] o 로컬 이웃 내부에서 양자화 인덱스의 부호는 0과 같지 않음.
- [0293] o 로컬 이웃 내부에서 0이 아닌 양자화 인덱스의 수. 이 숫자는 최대값으로 잘릴 수 있다.
- [0294] o 로컬 이웃에서 양자화 인덱스의 절대값의 합. 이 숫자는 최대값으로 잘릴 수 있다.
- [0295] o 로컬 이웃에서 양자화 인덱스의 절대값의 합과 로컬 이웃 내에서 0이 아닌 양자화 인덱스의 수의 차이. 이 숫자는 최대값으로 잘릴 수 있다.
- [0296] 다시 말해서, 본 발명에 따른 실시예는, 예를 들어, 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 특성에 따라, 확률 모델의 하위 세트 중에서 현재 신경망 매개 변수에 대한 확률 모델을 선택하도록 구성된, 신경망 매개변수를 인코딩하기 위한 장치를 포함하고, 이 특성은,
- [0297] 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 0이 아닌 양자화 인덱스의 부호,
- [0298] 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련되고 0이 아닌, 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 수,
- [0299] 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 절대값의 합,
- [0300] 현재 신경망 매개변수가 관련된 부분에 인접하는 신경망의 부분과 관련된 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 절대값의 합과
- [0301] 현재 신경망 매개변수가 관련된 부분에 인접하는 신경망의 부분과 관련되고 0이 아닌, 이전에 인코딩된 신경망 매개변수의 양자화 인덱스의 수 간의
- [0302] 차이
- [0303] 중 하나 이상을 포함한다.
- [0304] 각각, 본 발명에 따른 실시예는, 예를 들어, 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 특성에 따라, 확률 모델의 하위 세트 중에서 현재 신경망 매개 변수에 대한 확률 모델을 선택하도록 구성된, 신경망 매개변수를 디코딩하기 위한 장치를 포함하고, 이 특성은,
- [0305] 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 0이 아닌 양자화 인덱스의 부호,
- [0306] 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련되고 0이 아닌, 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 수,
- [0307] 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 절대값의 합,

- [0308] 현재 신경망 매개변수가 관련된 부분에 인접하는 신경망의 부분과 관련된 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 절대값의 합과
- [0309] 현재 신경망 매개변수가 관련된 부분에 인접하는 신경망의 부분과 관련되고 0이 아닌, 이전에 디코딩된 신경망 매개변수의 양자화 인덱스의 수 간의
- [0310] 차이
- [0311] 중 하나 이상을 포함한다.
- [0312] * 양자화 인덱스의 절대값에 대한 이진화는 양자화 인덱스의 절대값이 X , 예를 들어, $\text{abs_level_greater_}X$ 보다 큰지 여부를 지정하는 적응적으로 코딩된 bin을 포함한다. 이러한 bin을 코딩하는 데 사용되는 확률 모델(컨텍스트라고 함)은 후보 확률 모델 세트 중에서 선택된다. 선택된 확률 모델은 양자화 세트(허용 가능한 복원 레벨 세트) 또는 현재 양자화 인덱스에 대한 상태 변수에 의해 결정될 뿐만 아니라, 또한 계층에 대해 이미 전송된 양자화 인덱스에 의해 결정된다. 앞서 언급한 템플릿을 사용한다. 바람직한 실시예에서, 양자화 세트(또는 상태 변수)은 사용 가능한 확률 모델의 하위 세트(컨텍스트 세트라고도 함)를 결정하고, 이미 코딩된 양자화 인덱스의 데이터는 예를 들어 이 하위 세트(컨텍스트 세트) 내에서 사용된 확률 모델을 결정하고, 즉, 결정하는 데 사용할 수 있다. 확률 모델을 선택하기 위해, 위에서 설명한 방법(양자화 인덱스가 0과 같지 않은지 여부를 지정하는 bin에 대해)을 사용할 수 있다.
- [0313] 또한, 본 발명에 따른 장치는 이전에 인코딩된 신경망 매개변수(13)는 현재 신경망 매개변수(13')와 동일한 신경망 계층과 관련되도록 이전에 인코딩된 신경망 매개변수(13)의 위치를 지정하도록 구성될 수 있다.
- [0314] 또한, 예를 들어 본 발명에 따른 신경망 매개변수를 인코딩하기 위한 장치는, 하나 이상의 이전에 인코딩된 신경망 매개변수가 현재 신경망 매개변수가 참조하는 뉴런 상호 연결(11)이 관련되는 뉴런(10c), 또는 상기 뉴런에 인접한 추가 뉴런에서 나오거나, 이로부터 이어지는, 뉴런 상호 연결과 관련되도록 하는 방식으로 이전에 인코딩된 신경망 매개변수 중 하나 이상을 찾도록 구성될 수 있다.
- [0315] 추가 실시예에 따른 장치는 양자화 인덱스의 이진화의 하나 이상의 선행 bin에 대해 이전에 인코딩된 신경망 매개변수에 의존하는 확률 모델을 사용함으로써 및 하나 이상의 선행 bin을 따르는 양자화 인덱스의 이진화의 동등 확률 바이패스 모드 접미사 bin을 사용함으로써, 이진 산술 코딩을 사용하여 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 데이터 스트림(14)으로 인코딩하도록 구성될 수 있다.
- [0316] 양자화 인덱스의 이진화의 접미사 bin은 양자화 인덱스의 값을 이진화하기 위한 접미사 이진화의 이진화 코드의 bin을 나타낼 수 있으며, 그 절대값은 하나 이상의 리딩 bin에 의해 표현 가능한 최대 절대값을 초과한다. 따라서, 본 발명의 실시예에 따른 장치는 이전에 인코딩된 신경망 매개변수(13)의 양자화 인덱스(56)에 따라 접미사 이진화를 선택하도록 구성될 수 있다.
- [0317] 각각, 예를 들어 본 발명에 대한 신경망 매개변수 디코딩을 위한 장치는 이전에 디코딩된 신경망 매개변수는 현재 신경망 매개변수(13')와 동일한 신경망 계층과 관련되도록 이전에 디코딩된 신경망 매개변수(13)의 위치를 지정하도록 구성될 수 있다.
- [0318] 추가 구현예에 따르면, 예를 들어, 본 발명에 따른 신경망 매개변수를 디코딩하기 위한 장치는, 하나 이상의 이전에 디코딩된 신경망 매개변수가, 현재 신경망 매개변수가 참조하는 뉴런 상호 연결과 관련된 뉴런(10c), 또는 상기 뉴런에 인접한 추가 뉴런에서 나오거나 이를 향하는, 뉴런 상호 연결(11)과 관련되도록 하는 방식으로 이전에 디코딩된 신경망 매개변수(13) 중 하나 이상을 찾도록 구성될 수 있다.
- [0319] 추가 실시예에 따른 장치는 양자화 인덱스의 이진화의 하나 이상의 선행 bin에 대해 이전에 디코딩된 신경망 매개변수에 의존하는 확률 모델을 사용함으로써 및 하나 이상의 선행 bin을 따르는 양자화 인덱스의 이진화의 동등 확률 바이패스 모드 접미사 bin을 사용함으로써, 이진 산술 코딩을 사용하여 데이터 스트림(14)으로부터 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩하도록 구성될 수 있다.
- [0320] 양자화 인덱스의 이진화의 접미사 bin은 양자화 인덱스의 값을 이진화하기 위한 접미사 이진화의 이진화 코드의 bin을 나타낼 수 있으며, 그 절대값은 하나 이상의 리딩 bin에 의해 표현 가능한 최대 절대값을 초과한다. 따라서, 실시예에 따른 장치는 이전에 디코딩된 신경망 매개변수의 양자화 인덱스에 따라 접미사 이진화를 선택하도록 구성될 수 있다.
- [0321] 4.5 인코딩 방법의 예

[0322] 왜곡(복원 품질)과 비트 전송률 간에 매우 양호한 균형을 제공하는 비트스트림을 얻기 위해서는, 양자화 인덱스는 라그랑주 비용 측정

$$D + \lambda \cdot R = \sum_k D_k + \lambda \cdot R_k = \sum_k \alpha_k \cdot (t_k - t'_k)^2 + \lambda \cdot R(q_k | q_{k-1}, q_{k-2}, \dots)$$

[0323]

[0324] 이 최소화되는 방식으로 선택되어야 한다. 독립 스칼라 양자화의 경우, 이러한 양자화 알고리즘(울-왜곡 최적화 양자화 또는 RDOQ라고 함)은 단락 2.1.1에서 논의되었다. 그러나 독립 스칼라 양자화와 비교할 때, 추가적인 어려움이 있다. 복원된 신경망 매개변수 t'_k 및 이에 따라 왜곡 $D_k = |t_k - t'_k|$ (또는 $D_{k,MSE} = (t_k - t'_k)^2$)은 연관된 양자화 인덱스 $q_k(56)$ 에만 의존하지 않고, 코딩 순서에서 선행 양자화 인덱스의 값에 대해서도 의존한다.

[0325] 그러나 단락 4.3.3에서 논의한 바와 같이, 신경망 매개변수(13) 간의 종속성은 격자 구조를 사용하여 표현될 수 있다. 추가 설명을 위해, 도 11에 제공된 바람직한 실시예를 예로 사용한다. 8개의 신경망 매개변수 세트의 예에 대한 격자 구조가 도 19에 나와 있다. 도 19는 본 발명의 실시예에 따른, (라그랑주 비용 측정 $D + \lambda \cdot R$ 과 같은) 비용 측정을 최소화하는 양자화 인덱스의 시퀀스 (또는 블록)를 결정하기 위해 이용될 수 있는 예시적인 격자 구조를 도시한다. 격자 구조는 4가지 상태를 갖는 종속 양자화의 바람직한 예를 나타낸다(도 18 참조). 격자는 8개의 신경망 매개변수(또는 양자화 인덱스)에 대해 표시된다. 제 1 상태(맨 왼쪽)는 초기 상태를 나타내며, 이는 0과 같다고 가정한다. 격자를 통한 경로(왼쪽에서 오른쪽으로)는 양자화 인덱스(56)에 대한 가능한 상태 전이를 나타낸다. 두 노드 간의 각 연결은 특정 하위 세트(A, B, C, D)의 양자화 인덱스를 나타낸다는 것에 유의한다. 각 부분 세트(A, B, C, D)에서 양자화 인덱스 $q_k(56)$ 을 선택하고 해당 울-왜곡 비용

$$J_k = D_k(q_k | q_{k-1}, q_{k-2}, \dots) + \lambda \cdot R_k(q_k | q_{k-1}, q_{k-2}, \dots)$$

[0326]

[0327] 을 두 격자 노드 간의 연관된 연결에 할당하면, 전체 울-왜곡 비용 $D + \lambda \cdot R$ 을 최소화하는 양자화 인덱스의 벡터/블록을 결정하는 문제는 격자를 통해 (도 19의 왼쪽에서 오른쪽으로) 최소 비용 경로를 갖는 경로를 찾는 것과 동일하다. 엔트로피 코딩에서 일부 종속성을 무시하면, 이 최소화 문제는 잘 알려진 비터비 알고리즘을 사용하여 해결할 수 있다.

[0328] 다시 말해서, 본 발명에 따른 실시예는 선택 및/또는 양자화를 수행하기 위해 비터비 알고리즘 및 울-왜곡 비용 측정을 사용하도록 구성된 장치를 포함한다.

[0329] 계층에 적합한 양자화 인덱스를 선택하기 위한 인코딩 알고리즘의 예는 다음과 같은 주요 단계로 구성될 수 있다:

[0330] 1. 초기 상태에 대한 울-왜곡 비용을 0으로 설정한다.

[0331] 2. 코딩 순서의 모든 신경망 매개변수(13)에 대해 다음을 수행한다.

[0332] a. 각 부분 세트 A, B, C, D에 대해, 주어진 원래 신경망 매개변수(13)에 대한 왜곡을 최소화하는 양자화 인덱스(56)를 결정한다.

[0333] b. 현재 신경망 매개변수(13')에 대한 모든 격자 노드(0, 1, 2, 3)에 대해 다음을 수행한다.

[0334] i. 이전 신경망 매개변수(13)의 상태를 현재 상태와 연결하는 두 경로에 대한 울-왜곡 비용을 계산한다. 비용은 선행 상태에 대한 비용과 $D_k + \lambda \cdot R_k$ 의 합으로 주어지고, 여기에서 D_k 및 R_k 는 고려된 연결과 관련된 하위 세트(A, B, C, D)의 양자화 인덱스를 선택하기 위한 왜곡 및 비율을 나타낸다.

[0335] ii. 계산된 비용의 최소값을 현재 노드에 할당하고 최소 비용 경로를 나타내지 않는 이전 신경망 매개변수(13)의 상태에 대한 연결을 제거한다.

[0336] 참고: 이 단계 후에, 현재 신경망 매개변수(13')에 대한 모든 노드는 이전 신경망 매개변수(13)에 대한 임의의 노드에 대한 단일 연결을 갖는다.

[0337] 3. 4개의 최종 노드의 비용(코딩 순서의 마지막 매개변수에 대해)을 비교하고 비용이 최소인 노드를 선택한다. 이 노드는 격자를 통과하는 고유한 경로와 연결되어 있다는 것에 유의한다(다른 모든 연결은 이전 단계에서 제거됨).

[0338] 4. 선택된 경로(최종 노드에 의해 지정됨)를 역순으로 따르고 격자 노드 사이의 연결과 관련된 양자화 인덱스

(56)를 수집한다.

[0339] 비터비 알고리즘에 기초한 양자화 인덱스(56)의 결정은 독립 스칼라 양자화를 위한 올-왜곡 최적화 양자화(RDOQ)보다 실질적으로 더 복잡하지 않다는 점에 유의해야 한다. 그럼에도 불구하고 종속 양자화를 위한 더 간단한 인코딩 알고리즘이 또한 있다. 예를 들어, 미리 정의된 초기 상태(또는 양자화 세트)로 시작하여, 양자화 인덱스(56)는 현재 양자화 인덱스의 영향만을 고려하는 임의의 비용 측정을 최소화함으로써 코딩/복원 순서로 결정될 수 있다. 현재 매개변수에 대해 결정된 양자화 인덱스(및 모든 선행 양자화 인덱스)가 주어지면, 다음 신경망 매개변수(13)에 대한 양자화 세트는 알려져 있다. 따라서 알고리즘은 코딩 순서에 따라 모든 신경망 매개변수에 적용할 수 있다.

[0340] 실시예에 따른 다음의 방법은 도 20, 21, 22 및 23에 도시되어 있다.

[0341] 도 20은 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 디코딩하기 위한 방법(400)의 블록도를 도시하고, 이 방법(400)은, 현재 신경망 매개변수에 대해, 이전 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스에 따라 복수의 복원 레벨 세트 중에서 복원 레벨 세트를 선택(54)함으로써, 데이터 스트림으로부터 현재 신경망 매개변수에 대한 양자화 인덱스를 디코딩(420)함으로써 - 자화 인덱스는 현재 신경망 매개변수에 대한 복원 레벨의 선택된 세트 중 하나의 복원 레벨을 나타냄 -, 및 현재 신경망 매개변수에 대한 양자화 인덱스에 의해 표시되는 복원 레벨의 선택된 세트의 하나의 복원 레벨로 현재 신경망 매개변수를 역양자화(62)함으로써, 신경망 매개변수들을 순차적으로 디코딩하는 단계를 포함한다.

[0342] 도 21은 데이터 스트림으로부터 신경망을 정의하는 신경망 매개변수를 인코딩하기 위한 방법(500)의 블록도를 도시하고, 이 방법(500)은 현재 신경망 매개변수에 대해, 이전에 인코딩된 신경망 매개변수에 대한 데이터 스트림으로 인코딩된 양자화 인덱스에 따라, 복수의 복원 레벨 세트 중에서 복원 레벨 세트를 선택(54)함으로써, 현재 신경망 매개변수를 복원 레벨의 선택된 세트의 하나의 복원 레벨로 양자화(64)함으로써, 및 현재 신경망 매개변수에 대한 양자화 인덱스가 데이터 스트림으로 양자화되는 하나의 복원 레벨을 나타내는 현재 신경망 매개변수에 대한 양자화 인덱스를 인코딩(530)함으로써, 신경망 매개변수들을 순차적으로 인코딩하는 단계를 포함한다.

[0343] 도 22는 본 발명의 실시예에 따라 신경망을 정의하는 신경망 매개변수를 복원하기 위한 방법의 블록도를 도시한다. 방법(600)은 신경망 매개변수 당, 제1 복원 계층 신경망 매개변수 값을 산출하기 위해 제1 복원 계층에 대한 제1 신경망 매개변수를 유도하는 단계(610)를 포함한다. 방법(600)은 신경망 매개변수당 제2 복원 계층 신경망 매개변수 값을 산출하기 위해 (예를 들어, 도 6에서 화살표(312)로 도시된 바와 같이), 데이터 스트림으로부터 제2 복원 계층에 대한 제2 신경망 매개변수를 디코딩하는 단계(620), 및 각각의 신경망 매개변수에 대해, 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써, (예를 들어, 도 6에서 화살표(314)로 도시된 바와 같이) 신경망 매개변수들을 복원하는 단계(630)를 더 포함한다.

[0344] 도 23은 본 발명의 실시예에 따른, 신경망을 정의하는 신경망 매개변수를 인코딩하는 방법의 블록도를 도시한다. 방법(700)은 신경망 매개변수당 제1 복원 계층 신경망 매개변수 값을 포함하는 제1 복원 계층에 대한 제1 신경망 매개변수를 사용하고, (예를 들어, 도 6에서 화살표(322)로 나타낸 바와 같이) 제2 복원 계층에 대한 제2 신경망 매개변수를 신경망 매개변수당 제2 복원 계층 신경망 매개변수 값을 포함하는, 데이터 스트림으로 인코딩하는 단계(710)를 포함하고, 여기서 신경망 매개변수는 각각의 신경망 매개변수에 대해, 제1 복원 계층 신경망 매개변수 값과 제2 복원 계층 신경망 매개변수 값을 결합함으로써 복원 가능하다.

[0345] 이하, 본 발명에 따른 추가적인 실시예가 제시될 것이다.

표 4

[0346]	quant_tensor(dimensions, maxNumNoRem, entryPointOffset) {	
	stateId = 0	997
	bitPointer = get_bit_pointer()	998
	lastOffset = 0	999
	for(i = 0; i < Prod(dimensions); i++) {	1000
	idx = TensorIndex(dimensions, i, scan_order)	1001
	if(entryPointOffset != -1 && GetEntryPointIdx(dimensions, i, scan_order) != -1) {	1002
	lvlCurrRange = 256	1003
	j = entryPointOffset + GetEntryPointIdx(dimensions, i, scan_order)	1004

lvlOffset = cabac_offset_list[j]	1005
if(dq_flag)	1006
stateId = dq_state_list[j]	1007
set_bit_pointer(bitPointer + lastOffset + BitOffsetList[j])	1008
lastOffset = BitOffsetList[j]	1009
Invoke initialisation process for probability estimation parameters	1010
}	1011
int_param(idx, maxNumNoRem, stateId)	1012
if(dq_flag) {	1013
nextSt = StateTransTab[stateId][QuantParam[idx] & 1]	1014
if(QuantParam[idx] != 0) {	1015
QuantParam[idx] = QuantParam[idx] << 1	1016
if(QuantParam[idx] < 0)	1017
QuantParam[idx] += stateId & 1	1018
else	1019
QuantParam[idx] += - (stateId & 1)	1020
}	1021
stateId = nextSt	1022
}	
}	
}	

[0347] 예를 들어 1014 행에 표시된, 2D 정수 배열 StateTransTab[][]은 종속 스칼라 양자화에 대한 상태 전이 표를 지정하며, 다음과 같다:

[0348] StateTransTab[][] = { {0, 2}, {7, 5}, {1, 3}, {6, 4}, {2, 0}, {5, 7}, {3, 1}, { 4, 6} }

표 5

[0349]

int_param(i, maxNumNoRem, stateId) {	
QuantParam[i] = 0	5997
sig_flag	5998
if(sig_flag) {	5999
QuantParam[i]++	6000
sign_flag	6001
j = -1	6002
do {	6003
j++	6004
abs_level_greater_x[j]	6005
QuantParam[i] += abs_level_greater_x[j]	6006
} while(abs_level_greater_x[j] == 1 && j < maxNumNoRem)	6007
if(j == maxNumNoRem) {	6008
RemBits = 0	6009
j = -1	6010
do {	6011
j++	6012
abs_level_greater_x2[j]	6013
if(abs_level_greater_x2[j]) {	6014
RemBits++	6015
QuantParam[i] += 1 << RemBits	6016
}	6017
} while(abs_level_greater_x2[j] && j < 30)	6018
abs_remainder	6019
QuantParam[i] += abs_remainder	6020
}	6021
QuantParam[i] = sign_flag ? -QuantParam[i] : QuantParam[i]	6022
}	

}	
---	--

- [0350] 이 프로세스에 대한 입력은 다음과 같다:
- [0351] * 디코딩할 텐서의 크기를 지정하는 변수 `tensorDims`.
- [0352] * 디코딩을 위한 진입점이 존재하는지 여부 및 진입점이 있는 경우 진입점 오프셋을 나타내는 변수 `entryPointOffset`.
- [0353] * 코드북 적용 여부, 및 코드북이 적용되는 경우 어떤 코드북이 사용되는지를 나타내는 변수 `codebookId`.
- [0354] 이 프로세스의 출력은 `tensorDims`와 같은 크기를 갖는 `TENSOR_FLOAT` 유형의 변수 `recParam`이다.
- [0355] 변수 `stepSize`는 다음과 같이 유도된다:
- [0356] 3001 $mul = (1 \ll QpDensity) + ((qp_value + QuantizationParameter) \& ((1 \ll QpDensity) - 1))$
- [0357] 3002 $shift = (qp_value + QuantizationParameter) \gg QpDensity$
- [0358] 3003 $stepSize = mul * 2^{shift - QpDensity}$
- [0359] 변수 `recParam`은 다음과 같이 업데이트된다.
- [0360] 4001 $recParam = recParam * stepSize$
- [0361] 위의 - 위의 계산에 따르면, `recParam`은 항상 이진 분수로 나타낼 수 있다.
- [0362] 구문 요소 `sig_flag`의 경우 사용될 컨텍스트 또는 확률 추정을 나타내는 `ctxInc`의 유도 프로세스에 관하여
- [0363] 이 프로세스에 대한 입력은 현재 `sig_flag`, 상태 값 `stateId` 및 관련 `sign_flag`(있는 경우) 전에 디코딩된 `sig_flag`이다. 현재 `sig_flag` 이전에 `sig_flag`가 디코딩되지 않은 경우 0으로 간주된다. 이전에 디코딩된 `sig_flag`와 관련된 `sign_flag`가 디코딩되지 않은 경우, 0으로 가정한다.
- [0364] 이 프로세스의 출력은 변수 `ctxInc`이다.
- [0365] 변수 `ctxInc`는 다음과 같이 유도된다:
- [0366] - `sig_flag`가 0이면, `ctxInc`는 `stateId*3`으로 설정된다.
- [0367] - 그렇지 않고, `sign_flag`가 0이면, `ctxInc`는 `stateId*3+1`로 설정된다.
- [0368] - 그렇지 않으면, `ctxInc`는 `stateId*3+2`로 설정된다.
- [0369] 위의 예는 신경망 매개변수(13)를 데이터 스트림(14)으로/로부터 데이터 스트림(14)으로 코딩/디코딩하는 개념을 보여주고, 여기서 신경망 매개변수(13)는 신경망(10)의 신경망 상호연결부(11)의 가중치, 예를 들어 가중치 텐서의 가중치와 관련될 수 있다. 신경망 매개변수(13)의 디코딩/코딩은 순차적으로 수행된다. 텐서의 크기 당 가중치 수의 곱만큼 많은 가중치를 갖는 텐서의 가중치를 순환하는 for-next 루프(1000)를 참조한다. 가중치는 미리 결정된 `TensorIndex(dimensions, i, scan_order)` 순서로 스캔된다. 현재 신경망 매개변수 `idx(13')`에 대해, 이전 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(58)에 기초하여 연속적으로 업데이트되는 양자화 상태 `stateId`에 따라, 1018 및 1020에서 두 개의 복원 레벨 세트(52) 중 복원 레벨 세트가 선택된다. 특히, 현재 신경망 매개변수 `idx`에 대한 양자화 인덱스는 1012에서 데이터 스트림으로부터 디코딩되고, 여기서 양자화 인덱스는 현재 신경망 매개변수(13')에 대한 복원 레벨의 선택된 세트 중 하나의 복원 레벨을 나타낸다. 2의 복원 레벨 세트는 1018 및 1020에서의 양자화 상태 인덱스에 따라, 1016에서의 복제에 이어, 1 더하기 또는 1 빼기로 정의된다. 이 때 1018과 1020에서, 현재 신경망 매개변수(13')는 현재 신경망 매개변수(13')에 대한 양자화 인덱스 `QuantParam[idx]`에 의해 표시되는 복원 레벨의 선택된 세트의 하나의 복원 레벨로 실제로 역양자화된다. 단계 크기 `stepSize`는 3001-3003에서 복원 레벨 세트를 매개변수화하는 데 사용된다. 이 미리 결정된 양자화 단계 크기 `stepSize`에 대한 정보는 구문 요소 `qp_value`를 통해 데이터 스트림에서 유도된다. 후자는 각각 전체 텐서 또는 전체 NN 계층 또는 전체 NN에 대한 데이터 스트림에 코딩될 수 있다. 즉, 신경망(10)은 하나 이상의 NN 계층(10a, 10b)를 포함할 수 있고, 각 NN 계층에 대해, 미리 결정된 양자화 단계 크기(QP)에 대한 정보는 데이터 스트림(14)으로부터 각각의 NN 계층에 대해 유도될 수 있고, 각 NN 계층에 대해, 복수의 복원 레벨 세트는 각각의 NN 계층에 속하는 신경망 매개변수(13)를 역양자화하는 데 사용되기 위

해 각각의 NN 계층에 대해 유도된 미리 결정된 양자화 단계 크기를 사용하여 매개변수화될 수 있다.

[0370] 1018과 1020에서 볼 수 있는 바와 같이, $stateId = 0$ 에 대해 설정된 제1 복원 레벨은 여기서 미리 결정된 양자화 단계 크기의 0 및 짝수 배수를 포함하고, $stateId = 1$ 에 대해 설정된 제2 복원 레벨은 미리 결정된 양자화 단계 크기(QP)의 0 및 홀수 배수를 포함한다. 각 신경망 매개변수(13)에 대해, 중간 정수 값 $QuantParam[idx](IV)$ 는 1015 내지 1021에서 각각의 신경망 매개변수(13)에 대해 설정된 선택된 복원 레벨 및 각 신경망 매개변수에 대한 엔트로피 디코딩된 양자화 인덱스 $QuantParam[idx]$ 에 따라 유도된 다음에, 각 신경망 매개변수에 대해, 각각의 신경망 매개변수에 대한 중간 값은 4001에서 각각의 신경망 매개변수에 대한 미리 결정된 양자화 단계 크기와 곱해진다.

[0371] 현재 신경망 매개변수(13')에 대해, 두 개의 복원 레벨 세트(예: 세트 0, 세트 1) 중 복원 레벨 세트의 선택은, 1014에 도시된 바와 같이 이전에 디코딩된 신경망 매개변수에 대해 데이터 스트림으로부터 디코딩된 양자화 인덱스의 LSB 부분에 따라 이루어지고, 여기서 전이 표는 $QuantParam[idx]$ 의 LSB에 따라 $stateId$ 에서 다음 양자화 상태 $nextSto$ 로 전환되므로 $statId$ 는 이미 디코딩된 양자화 인덱스(56)의 과거 시퀀스에 의존한다. 따라서, 상태 전이는 이전에 디코딩된 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스(56)의 이전 함수의 결과, 즉 그 패리티에 의존한다. 다시 말해서, 현재 신경망 매개변수에 대해, 복수의 복원 레벨 세트 중에서 복원 레벨 세트의 선택은, 현재 신경망 매개변수에 대해 1018 및 1020에서 현재 신경망 매개변수와 연관된 상태 $statId$ 에 따라 복수의 복원 레벨 세트 중 복원 레벨 세트를 결정함으로써, 및 직전 신경망 매개변수, 즉 지금까지 $stateId$ 가 결정된 매개변수에 대해 데이터 스트림에서 디코딩된 양자화 인덱스에 따라, 1014에서 후속 신경망 매개변수, 반드시 다음에 코딩/디코딩될 NN 매개변수가 아니고, $stateId$ 가 다음에 결정되는 것에 대해 상태 $statId$ 를 업데이트함으로써, 상태 전이 프로세스에 의해 수행된다. 예를 들어, 여기에서 현재 신경망 매개변수는 다음에 코딩/디코딩될 NN 매개변수에 대한 $stateId$ 를 산출하기 위한 업데이트에 사용된다. 1014에서의 업데이트는 바로 앞의 (현재) 신경망 매개변수에 대한 데이터 스트림으로부터 디코딩된 양자화 인덱스의 이전 함수, 즉 그것의 패리티를 사용하여 수행된다. 상태 전환 프로세스는 8가지 가능한 상태 사이에서 전환하도록 구성된다. 전이는 $StateTransTab[][]$ 표를 통해 수행된다. 상태 전이 과정에서, 전이는 이 8가지 가능한 상태 사이에서 수행되며, 여기에서 1018 및 1020에서, 현재 신경망 매개변수에 대해 현재 신경망 매개변수와 연관된 상태 $stateId$ 에 따라 양자화 세트 중 복원 레벨 세트를 결정하는 단계는, 전반부의 가능한 상태, 즉 홀수 상태에 속하는 경우 2개의 복원 레벨 세트 중 제 1 복원 레벨 세트, 및 상태가 짝수의 가능한 상태, 즉 짝수 상태에 속하는 경우 2개의 복원 레벨 세트 중 제2 복원 레벨을 결정한다. 상태 $statId$ 의 업데이트는 바로 앞의 (현재) 신경망 매개변수에 대한 데이터 스트림에서 디코딩되는, 상태 $statId$ 와 양자화 인덱스(58)의 패리티, $QuantParam[idx] \& 1$ 을 후속 신경망 매개변수와 관련된 추가 상태에 매핑하는, $StateTransTab[][]$ 전이 표를 통해 수행된다.

[0372] 현재 신경망 매개변수에 대한 양자화 인덱스는 현재 신경망 매개변수에 대해 선택된 복원 레벨 세트 또는 보다 정확하게는 양자화 상태 $stateId$, 즉 현재 신경망 매개변수(13')에 대한 상태에 의존하는, 확률 모델을 사용하여 산술 코딩을 사용하여 데이터스트림으로 코딩되고, 데이터 스트림으로부터 디코딩된다. 1012에서 int_param 함수를 호출할 때 제 3 매개변수를 참조한다. 특히, 현재 신경망 매개변수에 대한 양자화 인덱스는 양자화 인덱스의 이전화의 적어도 하나의 빈, 여기서는 이전화 sig_flag , $sign_flag$ (선택 사항), $abs_level_greater_x[j]$, $abs_level_greater_x2[j]$ 및 $abs_remainder$ 중에서 빈 sig_flag 에 대한 현재 신경망 매개변수에 대한 상태에 의존하는 확률 모델을 사용함으로써, 이전 산술 코딩/디코딩을 사용하여 데이터 스트림으로 코딩 및 디코딩될 수 있다. sig_flag 는 현재 신경망 매개변수의 양자화 인덱스(56)가 0인지 아닌지를 나타내는 유의성 빈이다. 확률 모델의 종속성은 종속성을 사용하여 신경망 매개변수에 대한 컨텍스트 세트에서 컨텍스트를 선택하는 것을 포함하고, 각 컨텍스트는 관련된 미리 결정된 확률 모델을 갖는다. 여기에서, sig_flag 에 대한 컨텍스트는 $ctxInc$ 를 컨텍스트 목록에서 컨텍스트를 인덱싱하기 위한 인덱스에 대한 증분자로서 사용함으로써 선택되고, 각 컨텍스트는 이전 확률 모델과 연관된다. 모델은 컨텍스트와 연관된 빈을 사용하여 업데이트될 수 있다. 즉, 각각의 컨텍스트를 사용하여 산술적으로 코딩된 양자화 인덱스를 기반으로 컨텍스트 각각과 관련된 미리 결정된 확률 모델이 업데이트될 수 있다.

[0373] sig_flag 에 대한 확률 모델은 이전에 디코딩된 신경망 매개변수의 양자화 인덱스, 즉 이전에 디코딩된 신경망 매개변수의 sig_flag 및 그 부호를 나타내는 그 $sign_flag$ 에 추가로 의존하는 것에 유의한다. 더 정확하게 말하면, 상태 $stateId$ 에 따라, 복수의 확률 모델 중에서, 즉 컨텍스트 증분기 상태 0~23 중에서, 확률 모델의 하위 세트, 즉, $\{0 \dots 23\}$ 중 3개의 연속적인 컨텍스트를 포함하는 8개가 미리 선택되고, sig_flag 에 대한 확률 모델의 하위 세트 중에서 현재 신경망 매개변수에 대한 확률 모델이 이전에 디코딩된 신경망 매개변수의 양자화 인덱스

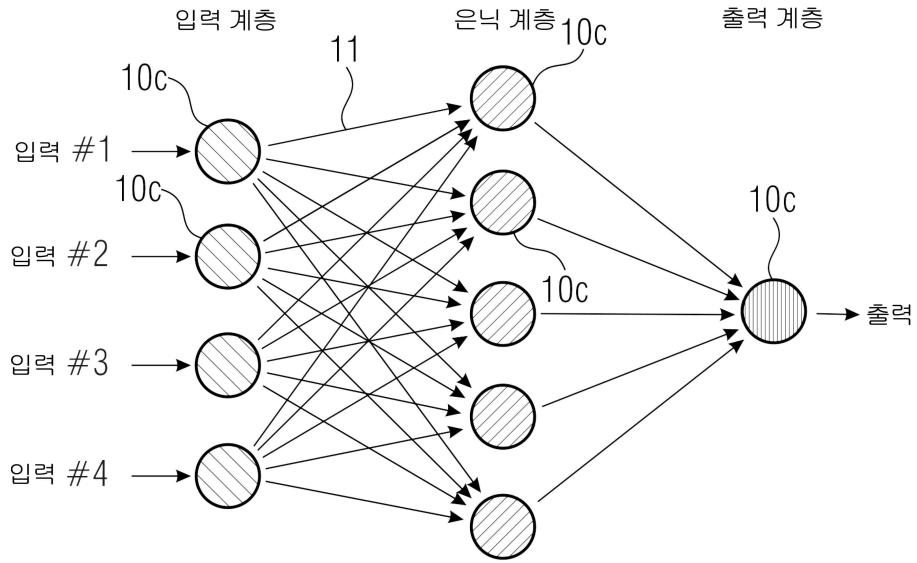
에 따라(121), 즉 이전 NN 매개변수의 sig_flag 및 sign_flag를 기반으로 하여 선택된다. stateID의 제 1 값에 대해 미리 선택된 하위 세트는 stateID의 다른 값에 대해 미리 선택된 하위 세트와 분리된다. sig_flag 및 sgin_flag가 사용되는 이전 NN 매개변수는 현재 신경망 매개변수가 관련된 부분에 인접한 신경망의 부분과 관련된다.

- [0374] 복수의 실시예가 위에서 설명되었다. 실시예의 측면 및 특징은 개별적으로 또는 조합하여 사용될 수 있다는 점에 유의해야 한다. 또한, 본 발명의 제1 및 제2 측면에 따른 실시예의 측면 및 특징이 조합되어 사용될 수 있다.
- [0375] 추가 실시예는 장치를 포함하고, 이 때 신경망 매개변수는 예를 들어 신경망(10)이 표현되는 복원 계층의 하나의 복원 층, 예를 들어, 향상 계층과 관련된다. 장치는 신경망이 신경망 매개변수 별로, 신경망 매개변수를, 대응하는, 예를 들어 공통 뉴런 상호 연결과 관련된 것, 또는 솔직히 말해서, 다른 표현 계층에서 NN 계층의 행렬 표현에 같은 위치에 있는 것들, 하나 이상의 추가 복원 계층의 신경망 매개변수와 결합하여 복원 가능하도록 구성될 수 있다.
- [0376] 예를 들어, 이 실시예와 함께 설명되는 바와 같이, 본 발명의 제1 및 제2 측면의 특징 및 측면이 조합될 수 있다. 제2 측면에 따른 종속항의 기능적 특징은 추가 실시예를 산출하기 위해 여기에 전이 가능하다.
- [0377] 또한, 본 발명의 측면에 따른 장치는 현재 신경망 매개변수에 해당하는 해당 신경망 매개변수에 의존하는 확률 모델을 사용하는 산술 인코딩을 사용하여 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 데이터 스트림(14)으로 인코딩하도록 구성될 수 있다.
- [0378] 제각기, 추가 실시예는 장치를 포함하고, 이 때 신경망 매개변수는 신경망(10)이 표현되는 복원 계층의 하나의 복원 계층, 예를 들어, 향상 계층에 관한 것이다. 장치는 신경망 매개변수 별로, 신경망 매개변수를 해당하는 예를 들어 공통 뉴런 상호 연결과 관련된 것, 또는 솔직히 말해서, 다른 표현 계층에서 NN 계층의 행렬 표현에 같은 위치에 있는 것들, 하나 이상의 추가 복원 계층의 신경망 매개변수와 결합하여 신경망을 복원하도록 구성될 수 있다.
- [0379] 예를 들어, 이 실시예와 함께 설명된 바와 같이, 본 발명의 제1 및 제2 측면의 특징 및 측면이 조합될 수 있다. 제2 측면에 따른 종속항의 기능적 특징은 추가 실시예를 산출하기 위해 여기에 양도될 수 있다.
- [0380] 또한, 본 발명의 측면에 따른 장치는 현재 신경망 매개변수에 해당하는 해당 신경망 매개변수에 의존하는 확률 모델을 사용하여 산술 코딩을 사용하여 데이터 스트림(14)으로부터 현재 신경망 매개변수(13')에 대한 양자화 인덱스(56)를 디코딩하도록 구성될 수 있다.
- [0381] 다시 말해, 복원 계층의 신경망 매개변수, 예를 들어 상술된 제 2 신경망 매개변수는 도 3 및 5 및 도 2 및 4에 대해 각각 설명된 개념에 따라 인코딩/디코딩 및/또는 양자화/역양자화될 수 있다.
- [0382] 일부 측면이 장치의 맥락에서 설명되었지만, 이들 측면은 또한 해당 방법에 대한 설명을 나타내며, 본 명세서에서 블록 또는 장치는 방법 단계 또는 방법 단계의 기능에 해당한다. 유사하게, 방법 단계의 맥락에서 설명된 측면은 또한 대응하는 블록 또는 대응하는 장치의 항목 또는 특징의 설명을 나타낸다.
- [0383] 본 발명의 데이터 스트림은 디지털 저장 매체에 저장될 수 있거나 무선 전송 매체 또는 인터넷과 같은 유선 전송 매체와 같은 전송 매체를 통해 전송될 수 있다.
- [0384] 특정 구현 요건에 따라, 본 발명의 실시예는 하드웨어 또는 소프트웨어로 구현될 수 있다. 구현은 각각의 방법이 수행되도록 프로그래밍 가능한 컴퓨터 시스템과 협력하는 (또는 협력할 수 있는), 전자적으로 판독 가능한 제어 신호가 저장된 플로피 디스크, DVD, Blu-Ray, CD, ROM, PROM, EPROM, EEPROM 또는 FLASH 메모리와 같은 디지털 저장 매체를 사용하여 수행할 수 있다.
- [0385] 본 발명에 따른 일부 실시예는 프로그램 가능한 컴퓨터 시스템과 협력하여 본 명세서에 설명된 방법 중 하나가 수행되도록 할 수 있는, 전자적으로 판독 가능한 제어 신호를 갖는 데이터 캐리어를 포함한다.
- [0386] 일반적으로, 본 발명의 실시예는 프로그램 코드를 갖는 컴퓨터 프로그램 제품으로서 구현될 수 있으며, 프로그램 코드는 컴퓨터 프로그램 제품이 컴퓨터 상에서 실행될 때 방법들 중 하나를 수행하기 위해 동작한다. 프로그램 코드는 예를 들어 기계 판독 가능한 캐리어에 저장될 수 있다.
- [0387] 다른 실시예는 기계 판독 가능한 캐리어에 저장된, 본 명세서에 설명된 방법 중 하나를 수행하기 위한 컴퓨터 프로그램을 포함한다.

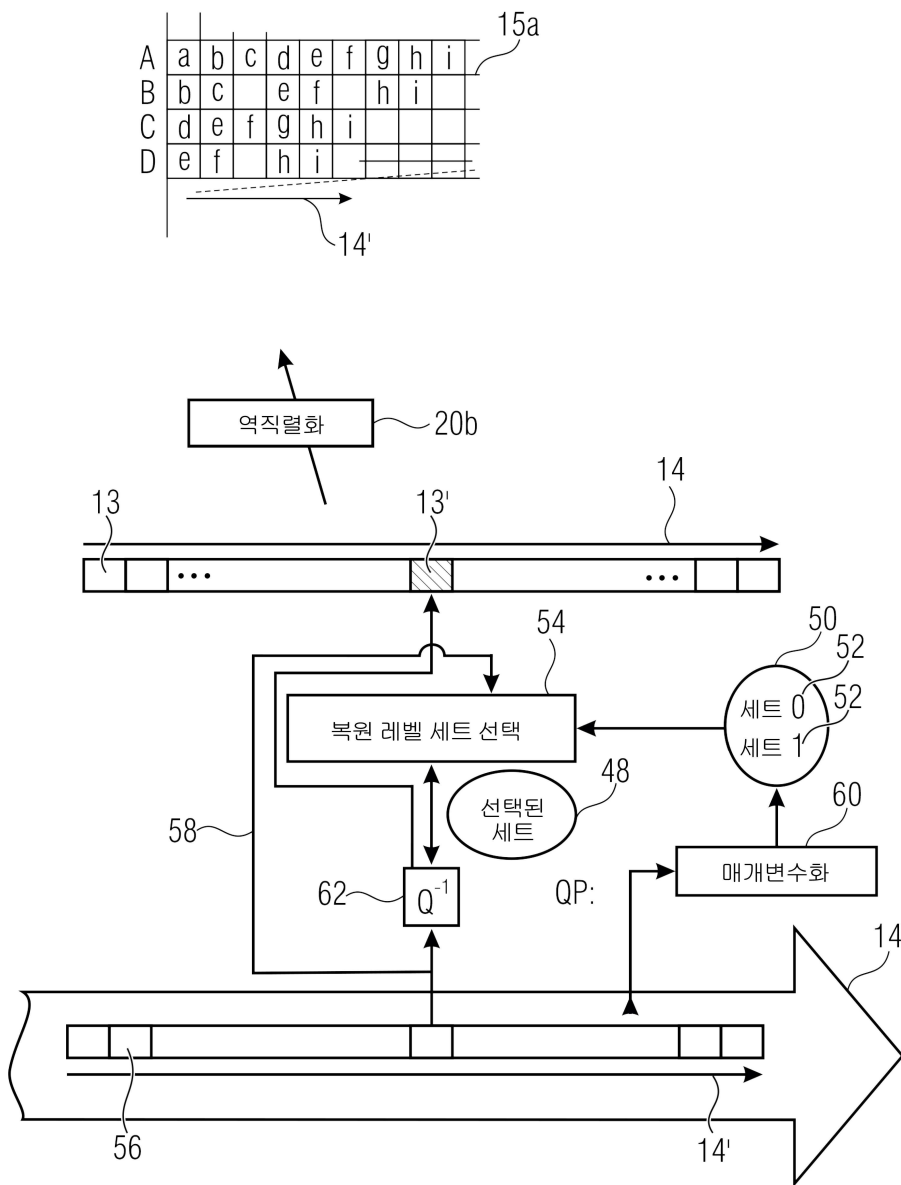
- [0388] 다시 말해, 본 발명의 방법의 실시에는 컴퓨터 프로그램이 컴퓨터에서 실행될 때, 본 명세서에서 설명된 방법들 중 하나를 수행하기 위한 프로그램 코드를 갖는 컴퓨터 프로그램이다.
- [0389] 따라서, 본 발명의 방법의 추가 실시에는 본 명세서에 기술된 방법 중 하나를 수행하기 위한 컴퓨터 프로그램을 포함하며 본 명세서에서 기록되어 있는, 데이터 캐리어(또는 디지털 저장 매체, 또는 컴퓨터 판독 가능 매체)이다.
- [0390] 따라서, 본 발명의 방법의 다른 실시에는 본 명세서에서 설명된 방법들 중 하나를 수행하기 위한 컴퓨터 프로그램을 나타내는 데이터 스트림 또는 신호들의 시퀀스이다. 데이터 스트림 또는 신호의 시퀀스는 예를 들어 인터넷을 통해 데이터 통신 연결을 통해 전송되도록 구성될 수 있다.
- [0391] 추가 실시에는 본 명세서에 설명된 방법들 중 하나를 수행하도록 구성되거나 적응된 처리 수단, 예를 들어 컴퓨터, 또는 프로그램 가능 논리 장치를 포함한다.
- [0392] 추가 실시에는 여기에 설명된 방법 중 하나를 수행하기 위한 컴퓨터 프로그램이 설치된 컴퓨터를 포함한다.
- [0393] 일부 실시 예에서, 프로그램 가능 논리 장치(예를 들어, 필드 프로그램 가능 게이트 어레이)는 본 명세서에서 설명된 방법의 기능 중 일부 또는 전부를 수행하는 데 사용될 수 있다. 일부 실시예에서, 필드 프로그램 가능 게이트 어레이는 본 명세서에서 설명된 방법 중 하나를 수행하기 위해 마이크로프로세서와 협력할 수 있다. 일반적으로, 방법은 바람직하게는 임의의 하드웨어 장치에 의해 수행된다.
- [0394] 상술한 실시예는 본 발명의 원리를 설명하기 위한 것일 뿐이다. 당업자라면 본 명세서에 기술된 배열 및 세부사항의 수정 및 변형이 명백할 것이다. 따라서, 본 명세서의 실시예의 설명에 의해 제시된 특정 세부사항이 아니라 계류중인 특허 청구범위의 범위에 의해서만 제한한다.
- [0395] 참고문헌
- [0396] [1] C. W. P. V. J. C. J. T. B. C. E. S. Sharan Chetlur, "cuDNN: 딥 러닝을 위한 효율적인 프리미티브", arXiv: 1410.0759, 2014
- [0397] [2] MPEG, "멀티미디어 콘텐츠 설명 및 분석을 위한 신경망 압축 작업 초안 2", ISO/IEC JTC1/SC29/WG11 문서, w18784, Geneva, 2019년 10월
- [0398] [3] D. Marpe, H. Schwarz 및 T. Wiegand, "H.264/AVC 비디오 압축 표준의 컨텍스트 기반 적응형 이진 산술 코딩", 비디오 기술을 위한 회로 및 시스템에 대한 IEEE 트랜잭션, Vol. 13, No. 7, pp. 620-636, 2003년 7월.
- [0399] [4] H. Kirchhofer, J. Stegemann, D. Marpe, H. Schwarz und T. Wiegand, "JVET-K0430-v3 - CE5-related: 상태 기반 확률 추정기", JVET, Ljubljana, 2018.
- [0400] [5] ITU - 국제 전기 통신 연합, "ITU-T H.265 고효율 비디오 코딩", 시리즈 H: 시청각 및 멀티미디어 시스템 - 시청각 서비스의 인프라 - 움직이는 비디오 코딩, 2015년 4월.
- [0401] [6] B. Bross, J. Chen 및 S. Liu, "JVET-M1001-v6 - 다목적 비디오 코딩(초안 4)", JVET, 마라케시, 2019년.

도면

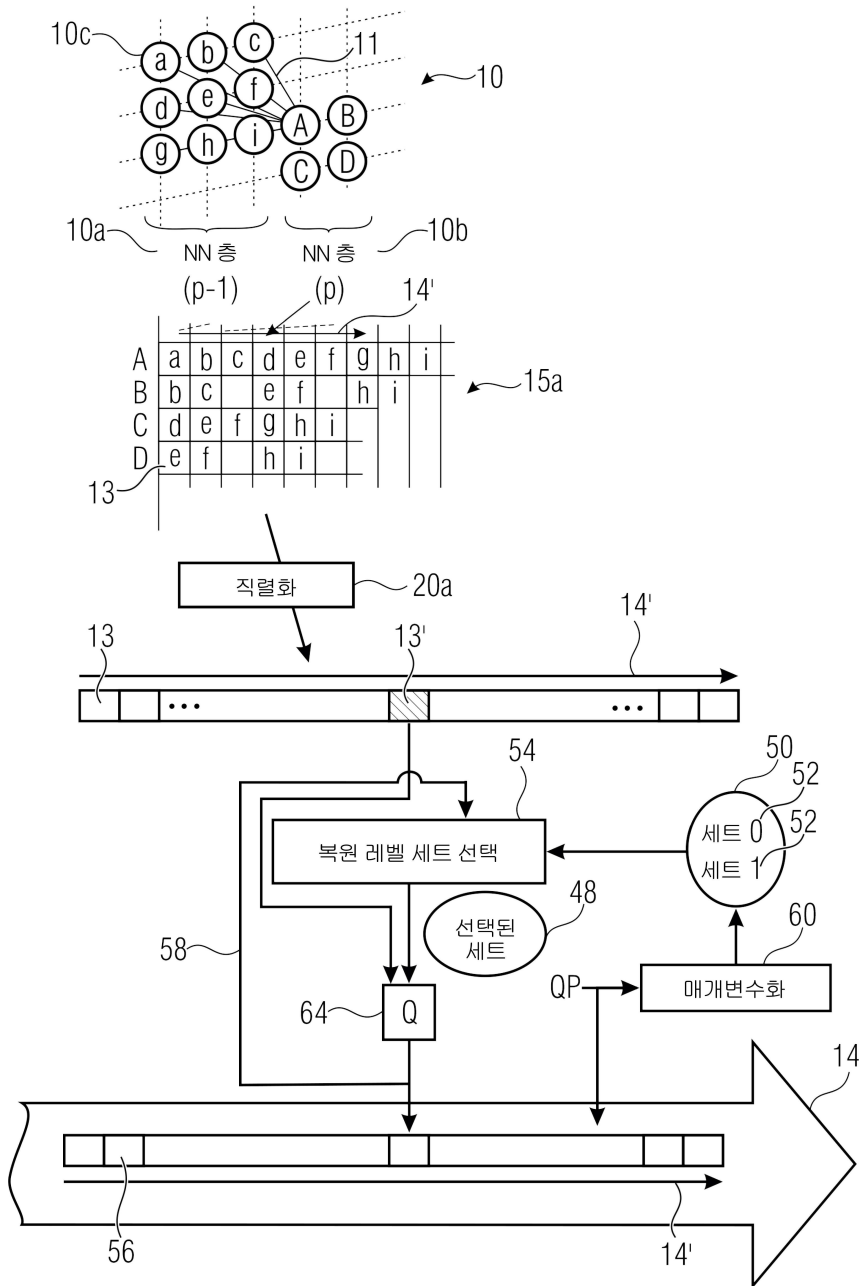
도면1



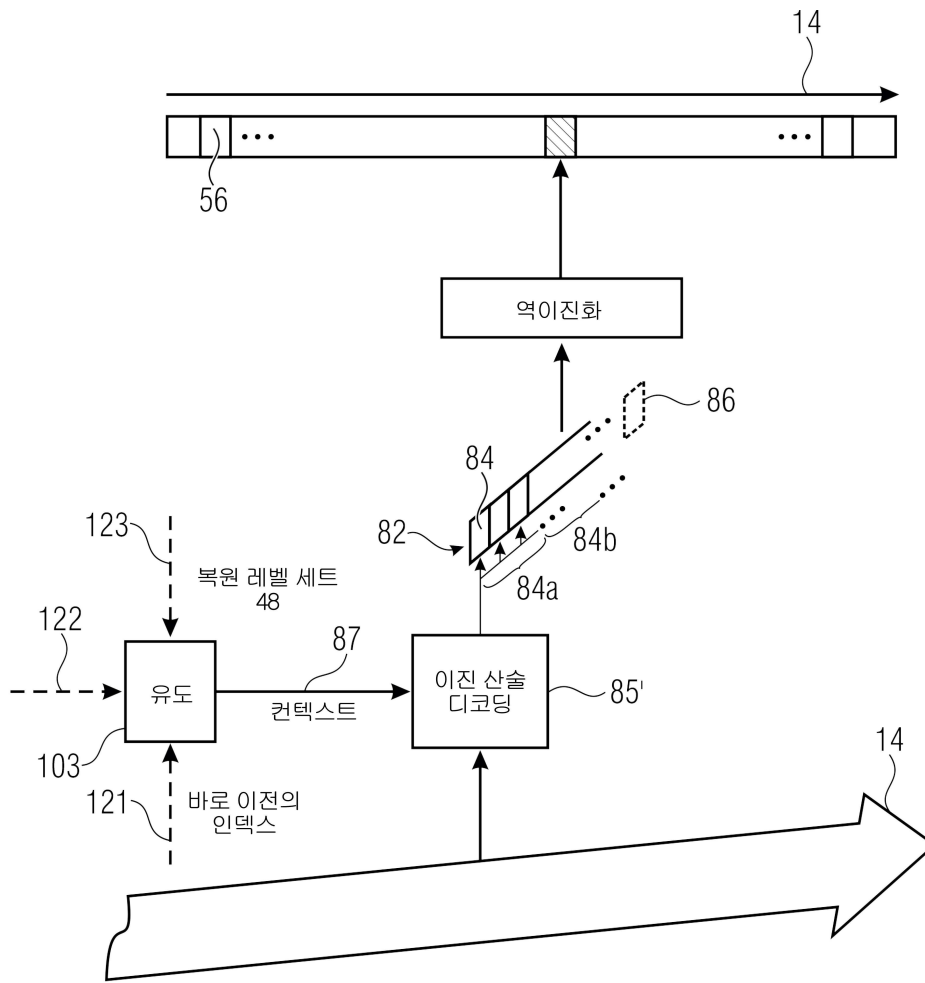
도면2



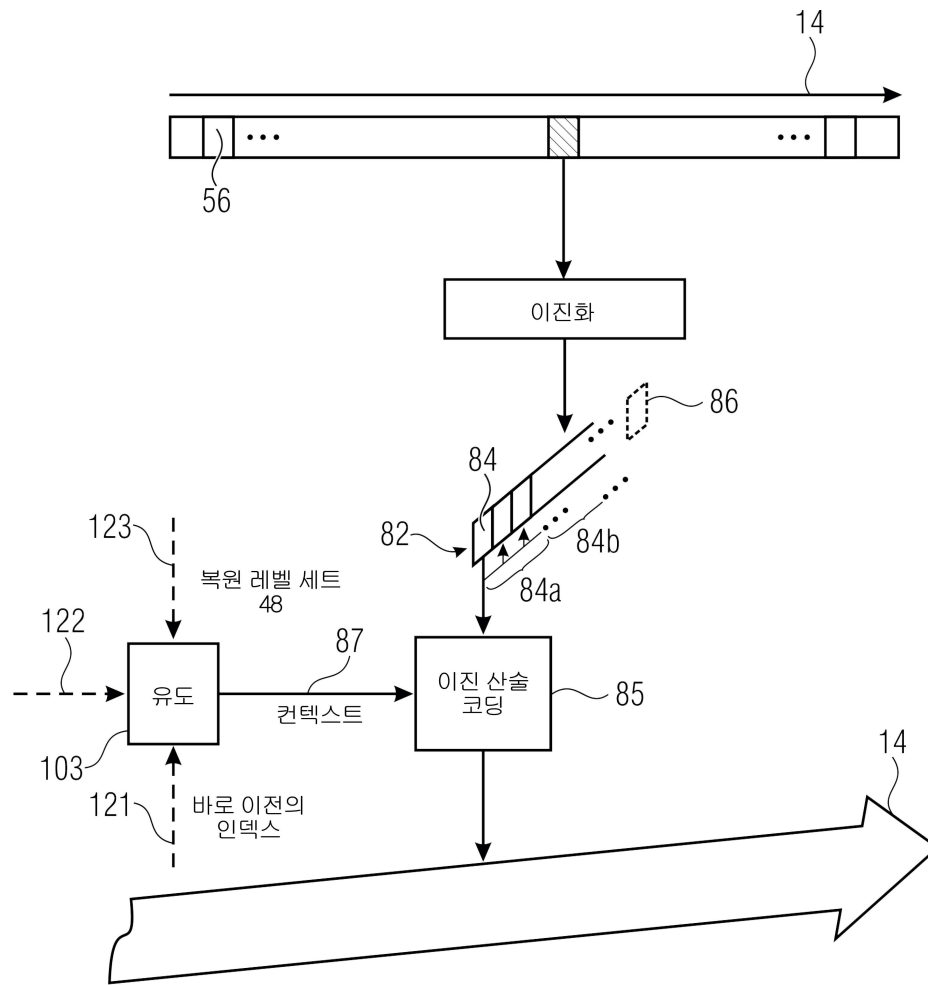
도면3



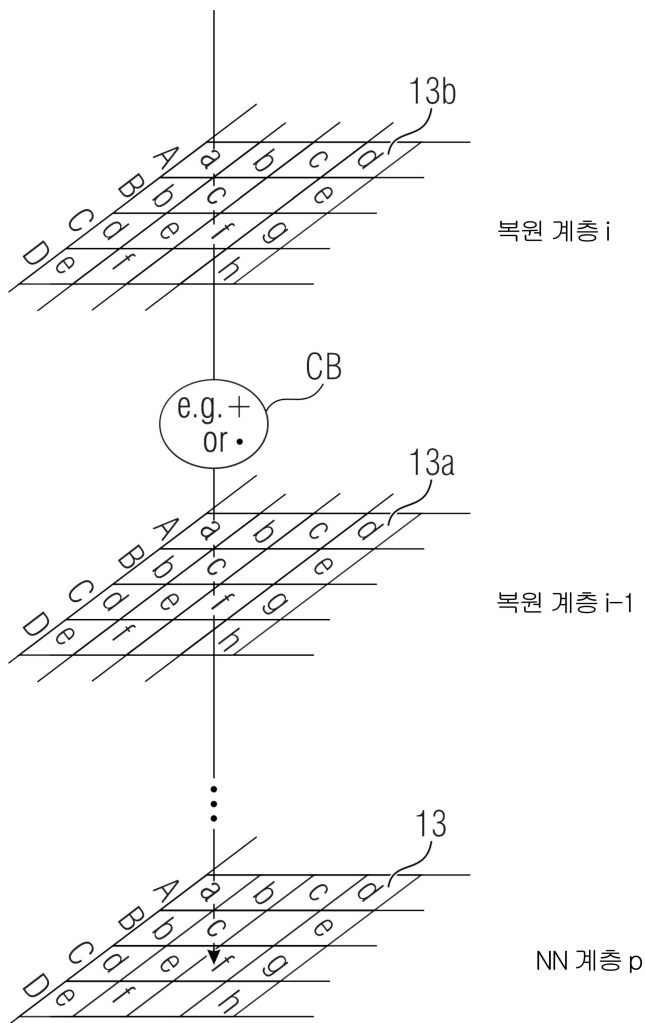
도면4



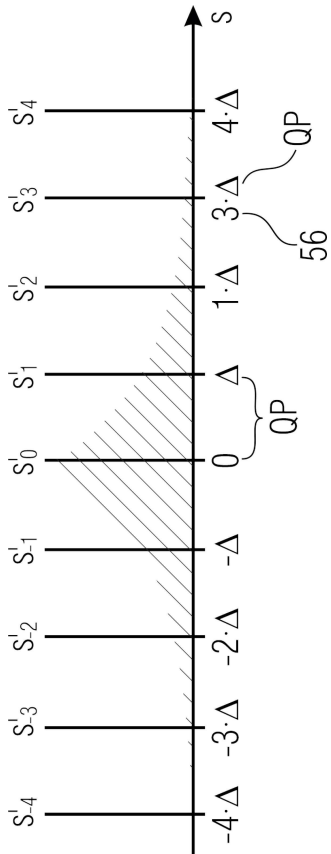
도면5



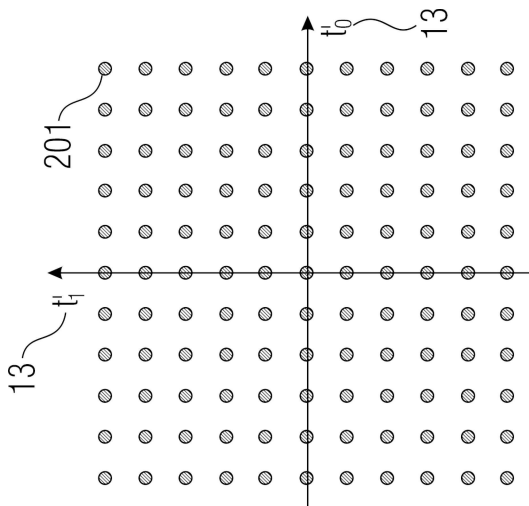
도면6



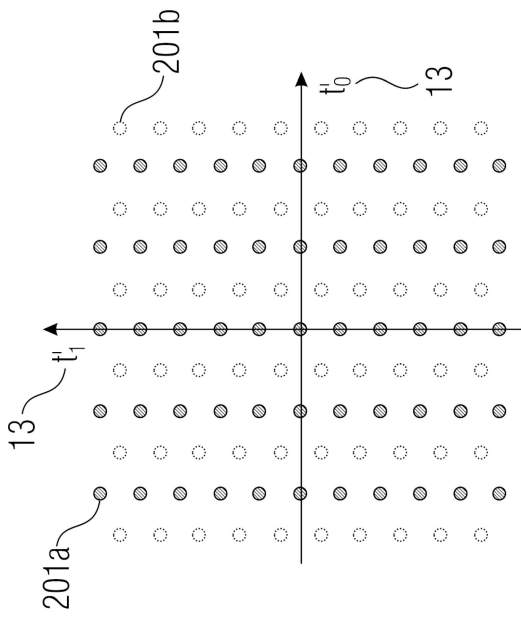
도면7



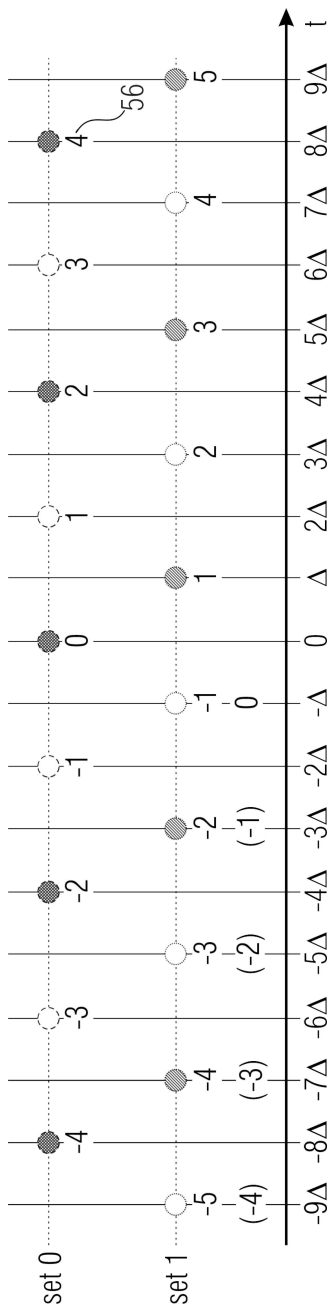
도면8a



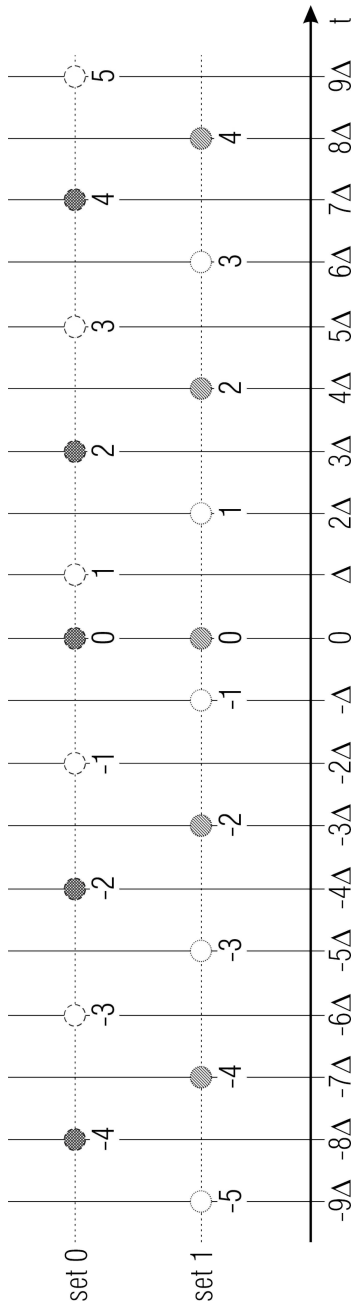
도면 8b



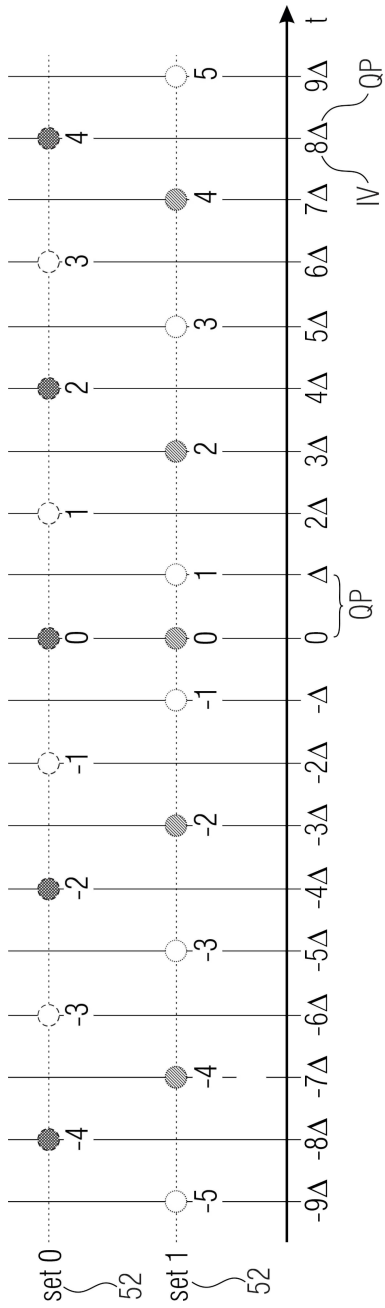
도면9a



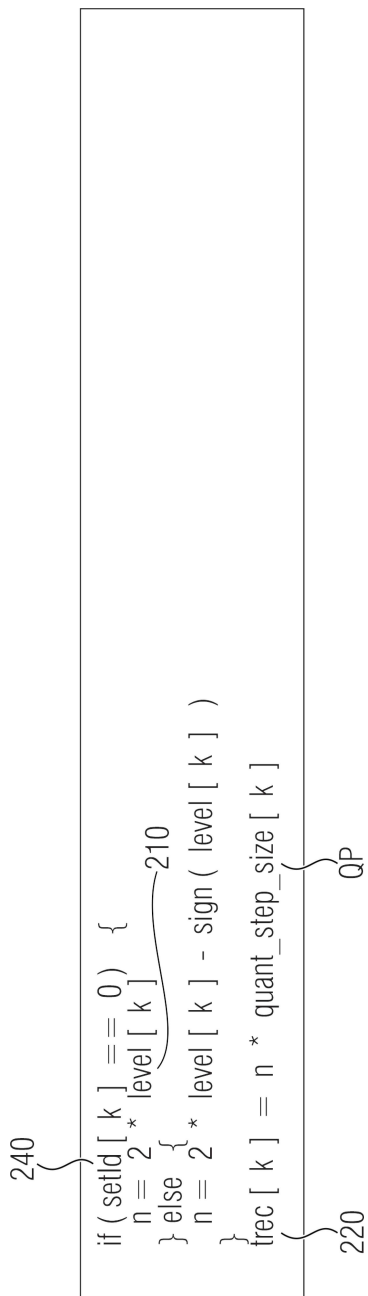
도면9b



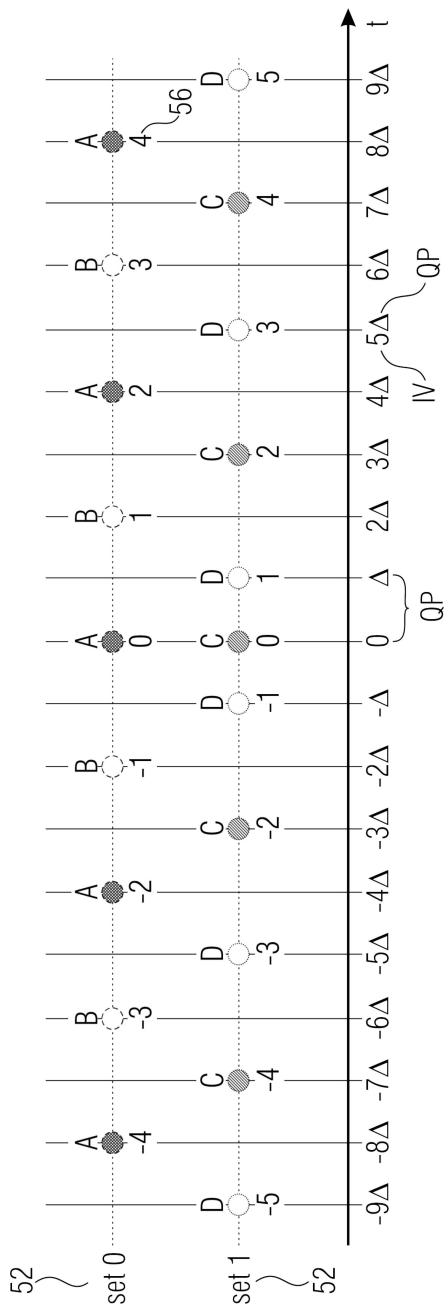
도면9c



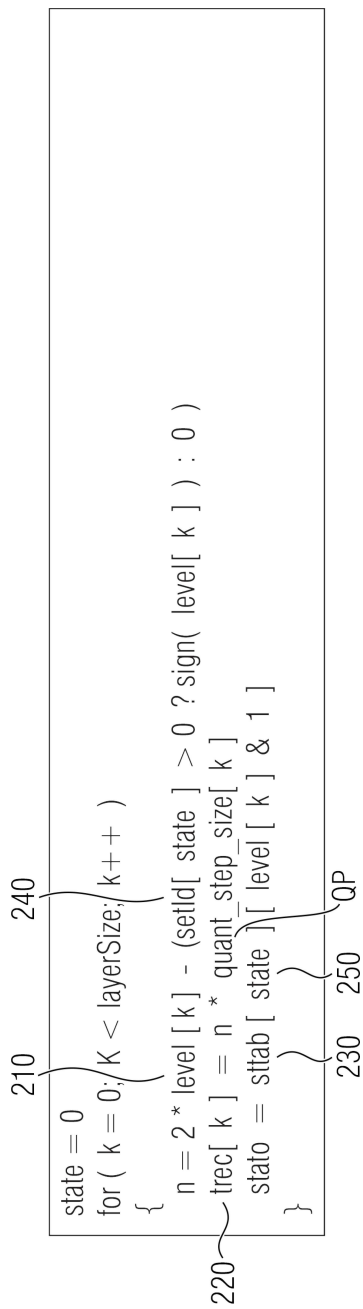
도면10



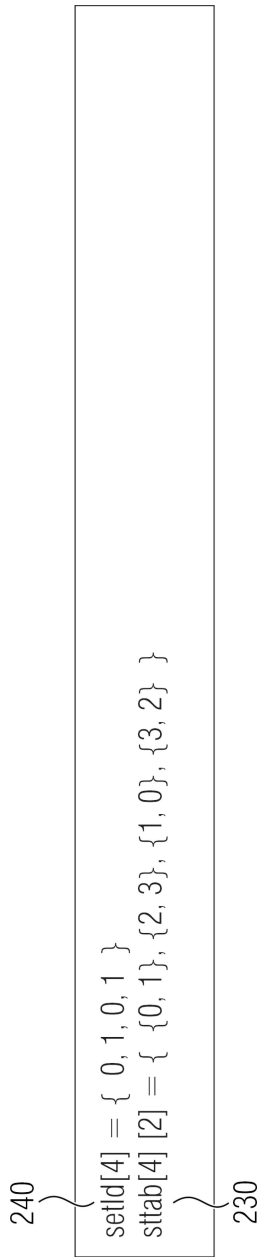
도면11



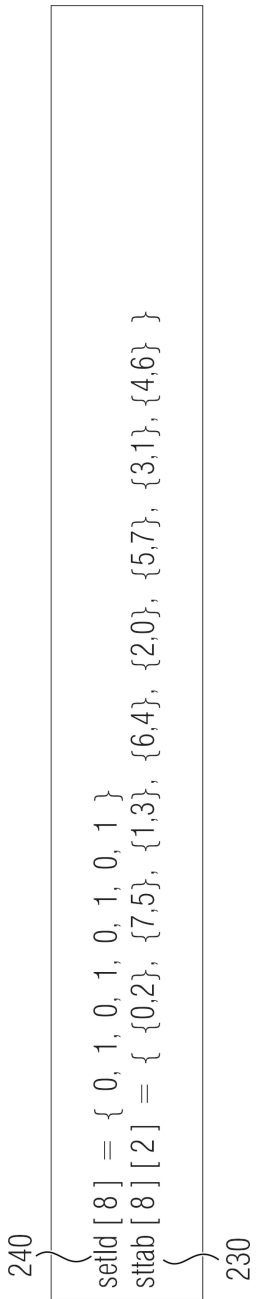
도면12



도면13



도면14



도면15

```

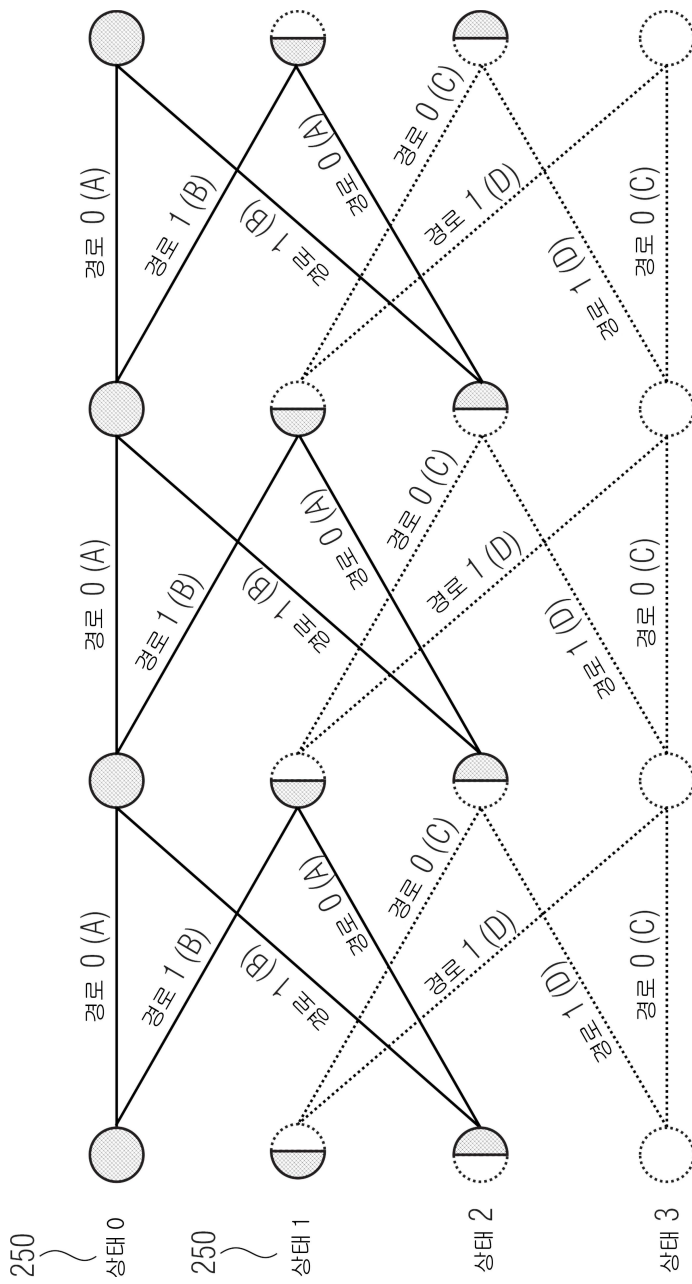
state = 0
for ( k = 0; k < layerSize; k++ )
{
    if ( level[ k ] != 0 ) QP 240 250
    {
        n = 2 * level[ k ] - ( setid[ state ] > 0 ? sign( level[ k ] ) : 0 )
        trec[ k ] = n * quant_step_size[ k ]
        state = sttab[ state ][ level[ k ] & 1 ]
    } else {
        trec[ k ] = 0 230
    }
}

```

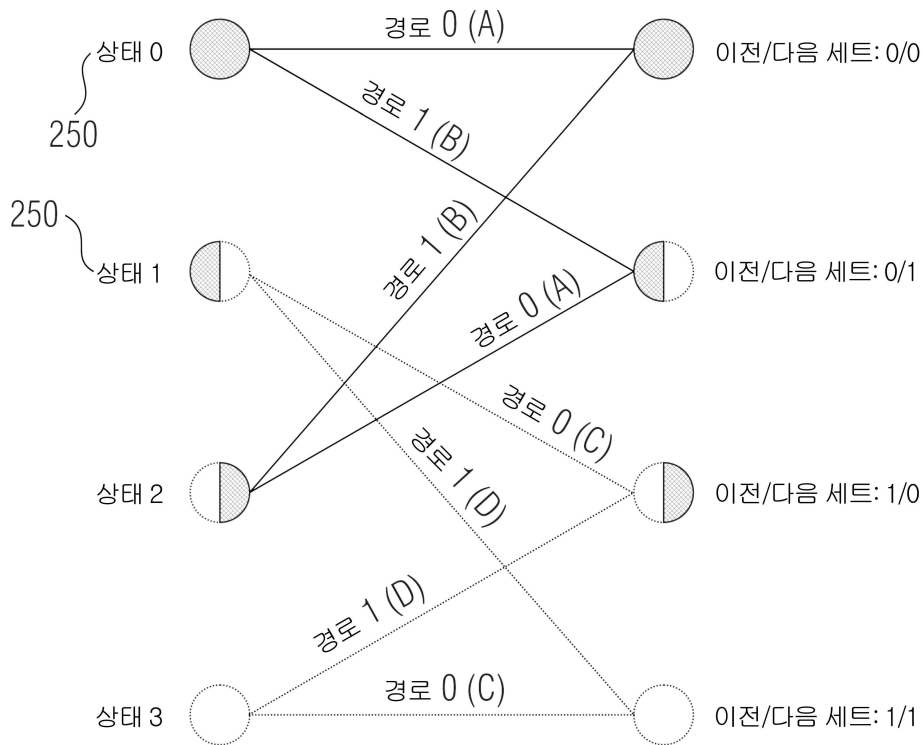
210

220

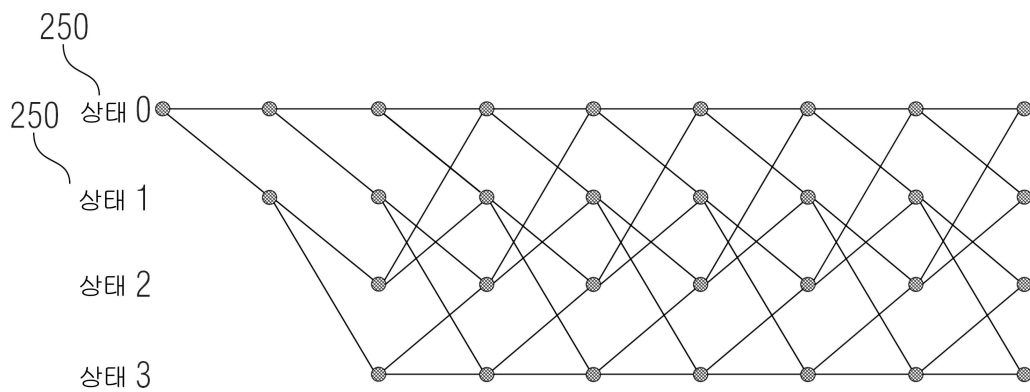
도면16



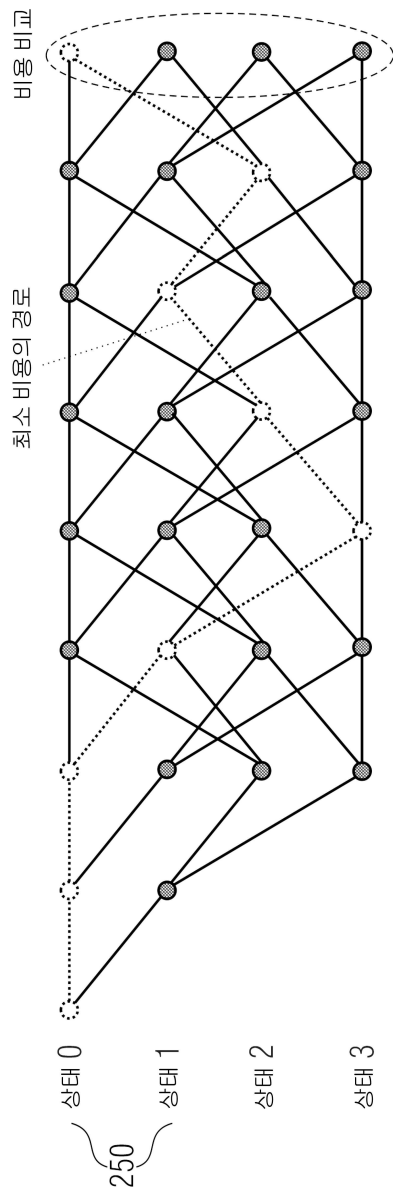
도면17



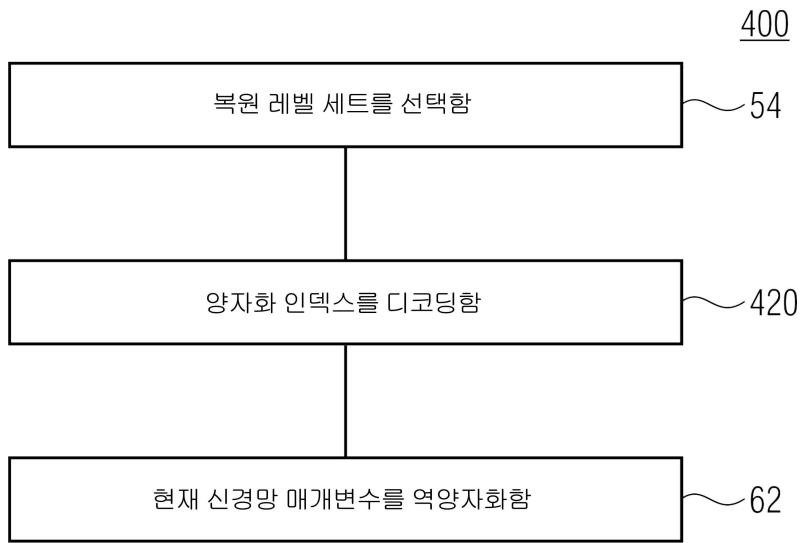
도면18



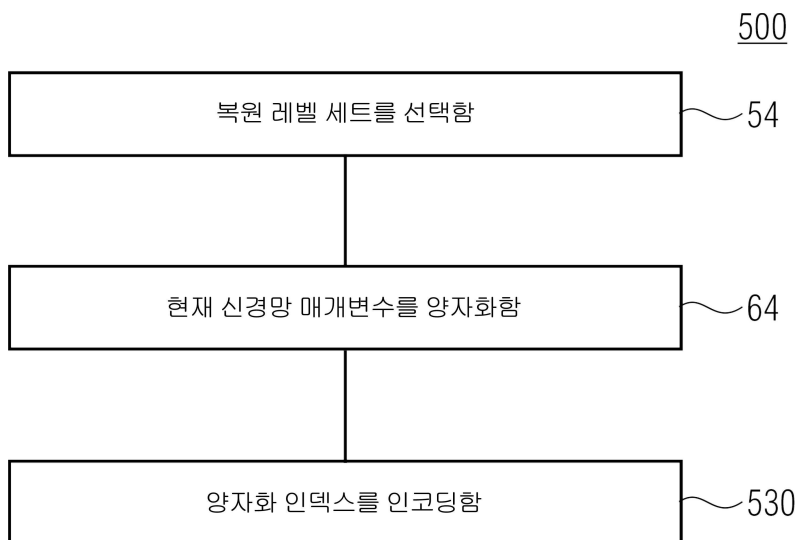
도면19



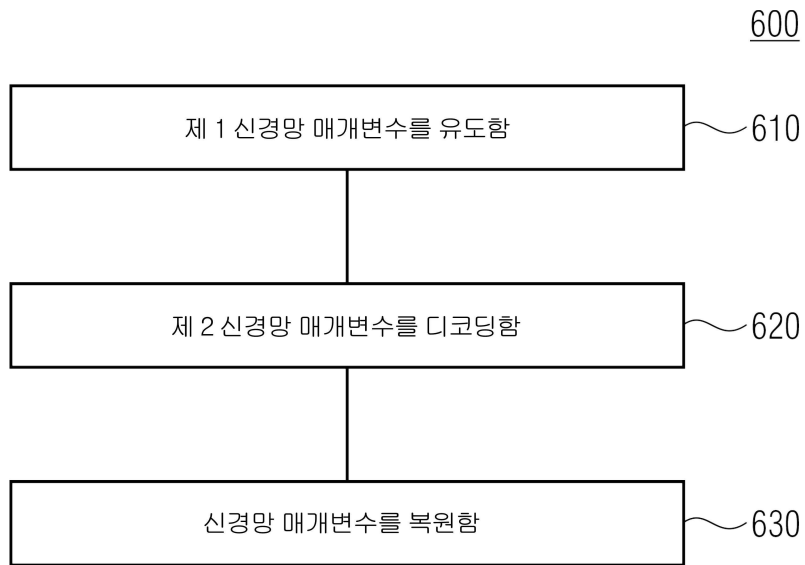
도면20



도면21



도면22



도면23

