(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년05월17일
(11) 등록번호 10-2533235
(24) 등록일자 2023년05월11일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 3/04* (2023.01)
(52) CPC특허분류
G06N 3/084 (2023.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2018-0132656
(22) 출원일자 2018년11월01일
심사청구일자 2021년06월17일
(65) 공개번호 10-2020-0055163
(43) 공개일자 2020년05월21일
(56) 선행기술조사문헌
US20170228639 A1*
Surat Teerapittayanon, BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks, arXiv:1709.01686v1, 1-7pages (2017. 9. 6.)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
이혁준
서울특별시 용산구 독서당로 111, 125동 1102호(한남동, 한남더힐)
소현
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(74) 대리인
정부연

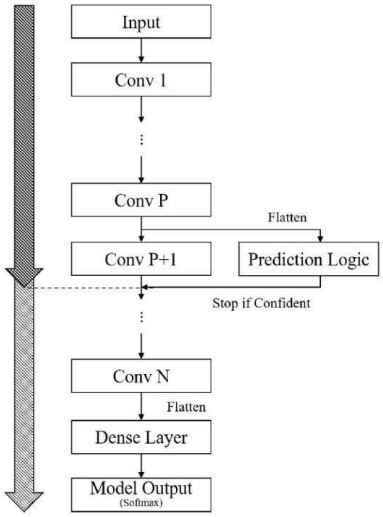
전체 청구항 수 : 총 12 항 심사관 : 양대경

(54) 발명의 명칭 **컨볼루션 신경망 기반의 입력 분류 장치 및 방법**

(57) 요 약

본 발명은 컨볼루션 신경망 기반의 입력 분류 장치 및 방법에 관한 것으로, 상호 연결되어 컨볼루션(convolution) 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결되어, 입력 데이터에 관한 분류를 미리 예측하는 예측 로직을 포함하는 컨볼루션 레이어부 및 상기 분류의 적정성을 결정하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 컨볼루션 연산 중단 결정부를 포함한다. 따라서, 본 발명은 전체 모델의 수행이 종료되는 시점을 앞당겨 추론에 필요한 수행 시간을 단축할 수 있다.

대 표 도 - 도4



명 세 서

청구범위

청구항 1

상호 연결되어 컨볼루션(convolution) 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결되어, 입력 데이터에 관한 분류를 미리 예측하는 예측 로직을 포함하는 컨볼루션 레이어부; 및

상기 분류의 적정성을 결정하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 컨볼루션 연산 중단 결정부를 포함하되,

상기 컨볼루션 레이어부는 상기 적어도 하나의 제1 컨볼루션 레이어에 있는 가장 마지막 컨볼루션 레이어의 출력을 상기 예측 로직에 제공하고, 상기 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 상기 예측 로직에 의한 예측 연산을 수행하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 2

삭제

청구항 3

제1항에 있어서, 상기 컨볼루션 레이어부는

상기 가장 마지막 컨볼루션 레이어와 직접적으로 연결된 상기 적어도 하나의 제2 컨볼루션 레이어에 있는 가장 최초 컨볼루션 레이어에 의한 컨볼루션 연산과 상기 예측 로직에 의한 예측 연산을 동시에 수행하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 4

삭제

청구항 5

제1항에 있어서, 상기 컨볼루션 레이어부는

상기 예측 로직의 개수가 복수인 경우에는 각각의 관점에서 상기 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 결정하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 6

제1항에 있어서, 상기 컨볼루션 연산 중단 결정부는

상기 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 동적으로 결정하여 상기 예측 로직의 연결성을 제어하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 7

제3항에 있어서, 상기 컨볼루션 연산 중단 결정부는

정규화된 상기 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 상기 분류의 적정성을 결정하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 8

제5항에 있어서, 상기 컨볼루션 연산 중단 결정부는

복수의 예측 로직들 각각에서 미리 예측된 분류의 적정성을 해당 컨볼루션 레이어에 의한 컨볼루션 연산과 동시에 수행하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 9

제8항에 있어서, 상기 컨볼루션 연산 중단 결정부는

상기 컨볼루션 연산 순서에 따라 복수의 예측 로직들에서 미리 예측된 분류의 적정성을 차례대로 결정하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 10

제3항에 있어서, 상기 컨볼루션 연산 중단 결정부는

상기 예측 연산에 따른 분류의 적정성을 결정하여 상기 가장 최초 컨볼루션 레이어와 직접 연결된 다음 컨볼루션 레이어를 통한 컨볼루션 연산의 지속적 수행 여부를 결정하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 장치.

청구항 11

컨볼루션 신경망 기반의 입력 분류 장치에서 수행되는 입력 분류 방법에 있어서,

(a) 컨볼루션(convolution) 레이어부에 포함되고, 상호 연결되어 컨볼루션 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결된 예측 로직에 의해 입력 데이터에 관한 분류를 미리 예측하는 단계; 및

(b) 컨볼루션 연산 중단 결정부에 의해 상기 분류의 적정성을 결정하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 단계를 포함하되,

상기 (a) 단계는 상기 적어도 하나의 제1 컨볼루션 레이어에 있는 가장 마지막 컨볼루션 레이어의 출력을 상기 예측 로직에 제공하는 단계를 포함하고, 상기 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 상기 예측 로직에 의한 예측 연산을 수행하는 단계를 포함하는 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 방법.

청구항 12

삭제

청구항 13

삭제

청구항 14

제11항에 있어서, 상기 (b) 단계는

정규화된 상기 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 상기 분류의 적정성을 결정하는 단계인 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 방법.

청구항 15

제11항에 있어서, 상기 (b) 단계는

복수의 예측 로직들 각각에서 미리 예측된 분류의 적정성을 해당 컨볼루션 레이어에 의한 컨볼루션 연산과 동시에 수행하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 단계인 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 방법.

청구항 16

제15항에 있어서, 상기 (b) 단계는

상기 컨볼루션 연산 순서에 따라 복수의 예측 로직들에서 미리 예측된 분류의 적정성을 차례대로 결정하는 단계인 것을 특징으로 하는 컨볼루션 신경망 기반의 입력 분류 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 컨볼루션 신경망 기반의 입력 분류 기술에 관한 것으로, 보다 상세하게는 전체 모델의 수행이 종료되는 시점을 앞당겨 추론에 필요한 수행 시간을 단축할 수 있는 컨볼루션 신경망 기반의 입력 분류 장치 및 방법에 관한 것이다.

배 경 기 술

[0003] 다양한 애플리케이션에 인공 신경망 기법을 적용하게 되면서 임베디드 환경에서도 해당 알고리즘을 사용하는 사례가 증가하고 있다. 일반적으로 정확도가 높은 신경망일수록 결과 도출까지 긴 시간이 필요하다. 그러나 임베디드 환경은 시간과 에너지에 제약 사항이 존재하며, 기존의 신경망은 제약 사항에 동적으로 대처하지 못하는 문제점을 가지고 있다.

[0004] 한국등록특허 제10-1298393(2013.08.13)호는 그래픽 처리 유닛 상에서 콘볼루션 신경망을 트레이닝하는방법에 관한 것으로, 콘볼루션 신경망은 그래픽 처리 유닛 상에서 일련의 텍스처로 구현되고, 망은 일련의 순방향 및 역방향 패스를 통하여 트레이닝되며, 오차 함수(error function)의 기울기에 따른 기울기 하강법(gradient descent method)에 의하여 각 역방향 패스 중에 수정된 바이어스 행렬 및 콘볼루션 커널을 포함한다.

[0005] 한국등록특허 제10-0820723(2008.04.02)호는 은닉노드 목표값을 가진 2개층 신경망을 이용한 분리 학습시스템 및 방법에 관한 것으로, 은닉노드의 목표값을 설정함으로써 네트워크의 구조 및 가중치를 갱신하는 규칙을 변경시키지 않고, 상층과 하층으로 분리하여 학습을 실행함으로써, 계산량을 줄여 신뢰성을 높일 수 있으며, 적은 용량의 저장 공간만을 요구하며 수렴이 빠르고 상대적으로 안정성이 보장됨으로써, 수렴의 확률을 높일 수 있는 효과를 제공할 수 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국등록특허 제10-1298393(2013.08.13)호

(특허문헌 0002) 한국등록특허 제10-0820723(2008.04.02)호

발명의 내용

해결하려는 과제

[0008] 본 발명의 일 실시예는 전체 모델의 수행이 종료되는 시점을 앞당겨 추론에 필요한 수행 시간을 단축할 수 있는 컨볼루션 신경망 기반의 입력 분류 장치 및 방법을 제공하고자 한다.

[0009] 본 발명의 일 실시예는 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 예측 로직에 의한 예측 연산을 수행하는 컨볼루션 신경망 기반의 입력 분류 장치 및 방법을 제공하고자 한다.

[0010] 본 발명의 일 실시예는 정규화된 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 분류의 적정성을 결정하여 컨볼루션 연산의 지속적 수행 여부를 결정할 수 있는 컨볼루션 신경망 기반의 입력 분류 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

[0012] 실시예들 중에서, 컨볼루션 신경망 기반의 입력 분류 장치는 상호 연결되어 컨볼루션(convolution) 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결되어, 입력 데이터에 관한 분류를 미리 예측하는 예측 로직을 포함하는 컨볼루션 레이어부 및 상기 분류의 적정성을 결정하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 컨볼루션 연산 중단 결정부를 포함한다.

[0013] 상기 컨볼루션 레이어부는 상기 적어도 하나의 제1 컨볼루션 레이어에 있는 가장 마지막 컨볼루션 레이어의 출력을 상기 예측 로직에 제공할 수 있다.

[0014] 상기 컨볼루션 레이어부는 상기 가장 마지막 컨볼루션 레이어와 직접적으로 연결된 상기 적어도 하나의 제2 컨볼루션 레이어에 있는 가장 최초 컨볼루션 레이어에 의한 컨볼루션 연산과 상기 예측 로직에 의한 예측 연산을 동시에 수행할 수 있다.

[0015] 상기 컨볼루션 레이어부는 상기 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 상기 예측 로직에 의한 예측 연산을 수행할 수 있다.

[0016] 상기 컨볼루션 레이어부는 상기 예측 로직의 개수가 복수인 경우에는 각각의 관점에서 상기 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 결정할 수 있다.

[0017] 상기 컨볼루션 연산 중단 결정부는 상기 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 동적으로 결정하여 상기 예측 로직의 연결성을 제어할 수 있다.

[0018] 상기 컨볼루션 연산 중단 결정부는 정규화된 상기 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 상기 분류의 적정성을 결정할 수 있다.

[0019] 상기 컨볼루션 연산 중단 결정부는 복수의 예측 로직들 각각에서 미리 예측된 분류의 적정성을 해당 컨볼루션 레이어에 의한 컨볼루션 연산과 동시에 수행하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있다.

[0020] 상기 컨볼루션 연산 중단 결정부는 상기 컨볼루션 연산 순서에 따라 복수의 예측 로직들에서 미리 예측된 분류의 적정성을 차례대로 결정할 수 있다.

[0021] 상기 컨볼루션 연산 중단 결정부는 상기 예측 연산에 따른 분류의 적정성을 결정하여 상기 가장 최초 컨볼루션 레이어와 직접 연결된 다음 컨볼루션 레이어를 통한 컨볼루션 연산의 지속적 수행 여부를 결정할 수 있다.

[0022] 실시예들 중에서, 컨볼루션 신경망 기반의 입력 분류 방법은 (a) 컨볼루션(convolution) 레이어부에 포함되고, 상호 연결되어 컨볼루션 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결된 예측 로직에 의해 입력 데이터에 관한 분류를 미리 예측하는 단계 및 (b) 컨볼루션 연산 중단 결정부에 의해 상기 분류의 적정성을 결정하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 단계를 포함한다.

[0023] 상기 (a) 단계는 상기 적어도 하나의 제1 컨볼루션 레이어에 있는 가장 마지막 컨볼루션 레이어의 출력을 상기 예측 로직에 제공하는 단계일 수 있다.

[0024] 상기 (a) 단계는 상기 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 상기 예측 로직에 의한 예측 연산을 수행하는 단계일 수 있다.

[0025] 상기 (b) 단계는 정규화된 상기 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 상기 분류의 적정성을 결정하는 단계일 수 있다.

[0026] 상기 (b) 단계는 복수의 예측 로직들 각각에서 미리 예측된 분류의 적정성을 해당 컨볼루션 레이어에 의한 컨볼

루션 연산과 동시에 수행하여 상기 컨볼루션 연산의 중단 또는 계속 여부를 결정하는 단계일 수 있다.

[0027] 상기 (b) 단계는 상기 컨볼루션 연산 순서에 따라 복수의 예측 로직들에서 미리 예측된 분류의 적정성을 차례대로 결정하는 단계일 수 있다.

발명의 효과

[0029] 개시된 기술은 다음의 효과를 가질 수 있다. 다만, 특정 실시예가 다음의 효과를 전부 포함하여야 한다거나 다음의 효과만을 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0030] 본 발명의 일 실시예에 따른 컨볼루션 신경망 기반의 입력 분류 장치 및 방법은 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 예측 로직에 의한 예측 연산을 수행할 수 있다.

[0031] 본 발명의 일 실시예에 따른 컨볼루션 신경망 기반의 입력 분류 장치 및 방법은 정규화된 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 분류의 적정성을 결정하여 컨볼루션 연산의 지속적 수행 여부를 결정할 수 있다.

도면의 간단한 설명

[0033] 도 1은 본 발명의 일 실시예에 따른 컨볼루션 신경망 기반의 입력 분류 시스템을 설명하는 도면이다.

도 2는 도 1에 있는 입력 분류 장치를 설명하는 블록도이다.

도 3은 도 1에 있는 입력 분류 장치에서 수행되는 입력 분류 과정을 설명하는 순서도이다.

도 4는 도 1에 있는 입력 분류 장치에서 수행되는 추론 수행의 일 실시예를 설명하는 예시도이다.

도 5는 도 2에 있는 컨볼루션 레이어부에 포함된 예측 로직에서 수행되는 예측 연산의 의사 코드를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0034] 본 발명에 관한 설명은 구조적 내지 기능적 설명을 위한 실시예에 불과하므로, 본 발명의 권리범위는 본문에 설명된 실시예에 의하여 제한되는 것으로 해석되어서는 아니 된다. 즉, 실시예는 다양한 변경이 가능하고 여러 가지 형태를 가질 수 있으므로 본 발명의 권리범위는 기술적 사상을 실현할 수 있는 균등물들을 포함하는 것으로 이해되어야 한다. 또한, 본 발명에서 제시된 목적 또는 효과는 특정 실시예가 이를 전부 포함하여야 한다거나 그러한 효과만을 포함하여야 한다는 의미는 아니므로, 본 발명의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.

[0035] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.

[0036] "제1", "제2" 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.

[0037] 어떤 구성요소가 다른 구성요소에 "연결되어"있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결될 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어"있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 한편, 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0038] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함하다"또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0039] 각 단계들에 있어 식별부호(예를 들어, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순

서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0040] 본 발명은 컴퓨터가 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있고, 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[0041] 여기서 사용되는 모든 용어들은 다르게 정의되지 않는 한, 본 발명이 속하는 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미를 지니는 것으로 해석될 수 없다.

[0043] 보조 분류기(Auxiliary Classifier)는 신경망의 중간에 소프트맥스(softmax)를 사용하여 분류기 역할을 수행하며, 전체 신경망에서 추가적으로 손실을 발생시키며 학습을 수행할 수 있다. 인공신경망은 마지막 계층의 오답을 역전파(back-propagation)하며 가중치를 조정하는 학습 단계를 수행할 수 있다. 그러나, 인공신경망은 신경망이 깊어질수록 초반부 계층으로 갈수록 기울기 소실(gradient vanishing) 문제가 발생할 수 있다. 이러한 경우, 보조 네트워크를 사용하게 되면 기울기 전파에 용이하며 동시에 정규화(regularization) 효과를 통해 심층 신경망의 정확도를 개선할 수 있다.

[0045] 도 1은 본 발명의 일 실시예에 따른 컨볼루션 신경망 기반의 입력 분류 시스템을 설명하는 도면이다.

[0046] 도 1을 참조하면, 컨볼루션 신경망 기반의 입력 분류 시스템(100)은 사용자 단말(110), 입력 분류 장치(130) 및 데이터베이스(150)를 포함할 수 있다.

[0047] 사용자 단말(110)은 분류 대상이 되는 입력 데이터를 제공하여 분류 결과를 확인할 수 있는 컴퓨팅 장치에 해당할 수 있다. 사용자 단말(110)은 스마트폰, 노트북 또는 컴퓨터로 구현될 수 있으며, 반드시 이에 한정되지 않고, 태블릿 PC 등 다양한 디바이스로도 구현될 수 있다. 사용자 단말(110)은 입력 분류 장치(130)와 네트워크를 통해 연결될 수 있고, 복수의 사용자 단말(110)들은 입력 분류 장치(130)와 동시에 연결될 수 있다.

[0048] 입력 분류 장치(130)는 추론을 위한 학습 모델을 생성할 수 있고 학습 모델을 이용하여 입력 데이터에 대한 분류를 수행하여 분류 결과를 제공할 수 있는 컴퓨터 또는 프로그램에 해당하는 서버로 구현될 수 있다. 입력 분류 장치(130)는 사용자 단말(110)과 유선 네트워크 또는 블루투스, WiFi 등과 같은 무선 네트워크로 연결될 수 있고, 유선 또는 무선 네트워크를 통해 사용자 단말(110)과 통신을 수행할 수 있다. 일 실시예에서, 입력 분류 장치(130)는 데이터베이스(150)와 연동하여 학습 모델을 생성하고 학습 모델을 기초로 입력 데이터에 대한 추론에 필요한 다양한 정보들을 저장할 수 있다. 한편, 입력 분류 장치(130)는 도 1과 달리, 데이터베이스(150)를 내부에 포함하여 구현될 수 있고, 사용자 단말(110)과 통합되어 하나의 장치로서 구현될 수 있다. 즉, 사용자 단말(110)은 자체적으로 독립적인 입력 분류 장치(130)로서의 역할을 수행할 수 있다.

[0049] 일 실시예에서, 입력 분류 장치(130)는 입력 데이터에 대한 분류를 수행하기 위한 전단계로서 분류를 위한 학습 모델을 구축하여 데이터베이스(150)에 저장할 수 있다. 여기에서, 분류를 위한 학습 모델은 컨볼루션 신경망(Convolution Neural Network, CNN)에 기초한 컨볼루션 신경망 모델에 해당할 수 있다.

[0050] 데이터베이스(150)는 입력 분류 장치(130)가 컨볼루션 신경망 학습을 수행하고 입력 데이터에 대한 분류 결과를 생성하는 과정에서 필요한 다양한 정보들을 저장할 수 있다. 예를 들어, 데이터베이스(150)는 사용자 단말(110)로부터 수신한 학습 데이터들을 저장할 수 있고, 컨볼루션 신경망을 기초로 학습을 통해 생성된 컨볼루션 신경망 모델에 관한 정보들을 저장할 수 있으며, 반드시 이에 한정되지 않고, 입력 데이터에 대한 분류 결과를 생성하는 과정에서 다양한 형태로 수집 또는 가공된 정보들을 저장할 수 있다.

[0052] 도 2는 도 1에 있는 입력 분류 장치를 설명하는 블록도이다.

[0053] 도2를 참조하면, 입력 분류 장치(130)는 컨볼루션 레이어부(210), 컨볼루션 연산 중단 결정부(230) 및 제어부(250)를 포함할 수 있다.

[0054] 컨볼루션 레이어부(210)는 상호 연결되어 컨볼루션(convolution) 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결되어, 입력 데이터에 관한 분류를 미리 예측하는 예측 로직을 포함할 수 있다. 컨볼루션(convolution) 연산은 컨볼루션 신경망(CNN)에서 수행되는 연산에 해당할 수 있고, 3차원 데이터의 공간

적 정보를 유지한 채 다음 레이어로 제공하는 역할을 수행할 수 있다. 보다 구체적으로, 컨볼루션 연산은 특정 높이 및 너비를 가지는 필터(Filter, Kernel)를 일정 간격(Stride)으로 이동해가며 입력 데이터에 적용하는 연산에 해당할 수 있다. 컨볼루션 연산은 하나의 필터를 사용할 수 있고, 복수의 필터들을 사용할 수 있다.

[0055] 컨볼루션 레이어(convolution layer)는 컨볼루션 신경망을 구성하는 레이어에 해당할 수 있고, 컨볼루션 연산이 수행될 수 있다. 또한, 예측 로직은 컨볼루션 레이어부(210)에 포함되어 구현될 수 있고, 특정 컨볼루션 연산에 대한 출력을 입력으로 하여 수행 중인 모델을 중단시킬지의 여부를 판단하면서 해당 지점에서의 분류 결과를 제공하는 역할을 수행할 수 있다. 예측 로직은 컨볼루션 신경망 모델에서 컨볼루션 레이어들 사이에 연결될 수 있고, 복수의 예측 로직들이 하나의 컨볼루션 신경망 모델에 연결되어 구성될 수 있다.

[0056] 일 실시예에서, 컨볼루션 레이어부(210)는 적어도 하나의 제1 컨볼루션 레이어에 있는 가장 마지막 컨볼루션 레이어의 출력을 예측 로직에 제공할 수 있다. 컨볼루션 레이어부(210)는 복수의 컨볼루션 레이어들을 포함할 수 있고, 컨볼루션 레이어 사이에 연결된 예측 로직을 기준으로 하여 적어도 하나의 제1 및 제2 컨볼루션 레이어들로 구분할 수 있다. 예를 들어, 총 10개의 컨볼루션 레이어들로 구성된 컨볼루션 레이어부(210)에 대해 위에서 아래의 방향으로 각 컨볼루션 레이어에 번호를 붙일 수 있다.

[0057] 만약 3번과 4번 컨볼루션 레이어 사이에 예측 로직이 연결된 경우 1번부터 3번까지의 컨볼루션 레이어들은 적어도 하나의 제1 컨볼루션 레이어에 해당할 수 있고, 4번부터 10번까지의 컨볼루션 레이어들은 적어도 하나의 제2 컨볼루션 레이어에 해당할 수 있다. 따라서, 적어도 하나의 제1 컨볼루션 레이어에 있는 가장 마지막 컨볼루션 레이어는 3번 컨볼루션 레이어에 해당할 수 있다. 컨볼루션 레이어부(210)는 3번 컨볼루션 레이어의 출력을 예측 로직에 제공할 수 있다.

[0058] 일 실시예에서, 컨볼루션 레이어부(210)는 가장 마지막 컨볼루션 레이어와 직접적으로 연결된 적어도 하나의 제2 컨볼루션 레이어에 있는 가장 최초 컨볼루션 레이어에 의한 컨볼루션 연산과 예측 로직에 의한 예측 연산을 동시에 수행할 수 있다. 총 10개의 컨볼루션 레이어들로 구성된 컨볼루션 레이어부(210)에 대해 위에서 아래의 방향으로 각 컨볼루션 레이어에 번호를 붙이고, 3번과 4번 컨볼루션 레이어 사이에 예측 로직이 연결된 경우 적어도 하나의 제2 컨볼루션 레이어에 있는 가장 최초 컨볼루션 레이어는 4번 컨볼루션 레이어에 해당할 수 있다.

[0059] 컨볼루션 레이어부(210)는 4번 컨볼루션 레이어에 의한 컨볼루션 연산과 예측 로직에 의한 예측 연산을 각각 동시에 수행할 수 있다. 즉, 3번 컨볼루션 레이어의 출력을 기초로 3번 컨볼루션 레이어에 직접적으로 연결된 4번 컨볼루션 레이어에 의한 컨볼루션 연산이 수행될 수 있고, 이와 병렬적으로 3번 컨볼루션 레이어의 출력을 기초로 예측 로직에 의한 예측 연산이 수행될 수 있다. 입력 분류 장치(130)는 컨볼루션 레이어에 의해 수행되는 컨볼루션 연산 결과와 예측 로직에 의해 수행되는 예측 연산 결과 중 어느 하나를 선택하여 다음 연산의 입력으로서 활용할 수 있다.

[0060] 일 실시예에서, 컨볼루션 레이어부(210)는 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 연산을 통해 출력벡터를 생성한 후 정규화함으로써 예측 로직에 의한 예측 연산을 수행할 수 있다. 여기에서, 가중치 벡터는 보조 분류기의 전연결망으로부터 가중치를 추출하여 일차원으로 저장한 벡터에 해당할 수 있다. 컨볼루션 레이어부(210)는 적어도 하나의 제1 컨볼루션 레이어의 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화할 수 있다.

[0061] 예를 들어, 컨볼루션 레이어의 출력이 2차원 행렬로 표현되는 경우 컨볼루션 레이어부(210)는 각 행 또는 각 열로 분리한 후 차례대로 연결하여 1차원 벡터로 표현할 수 다. 컨볼루션 레이어부(210)는 가장 마지막 컨볼루션 레이어의 출력을 1차원 벡터화하고 가중치 벡터와의 점곱(dot product) 연산을 통해 출력벡터를 생성할 수 있다. 출력벡터는 예측 로직에 의한 예측 연산의 출력을 산출하기 위한 중간값에 해당할 수 있다. 컨볼루션 레이어부(210)는 출력벡터를 정규화하여 예측 로직에 의한 예측 연산의 출력으로 결정할 수 있다. 컨볼루션 레이어부(210)는 출력벡터에 대한 정규화(normalize)를 통해 예측 로직에 의해 발생하는 모든 출력값에 대해 동일한 연산을 적용할 수 있다.

[0062] 일 실시예에서, 컨볼루션 레이어부(210)는 예측 로직의 개수가 복수인 경우에는 각각의 관점에서 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 결정할 수 있다. 적어도 하나의 제1 및 제2 컨볼루션 레이어들은 예측 로직이 연결된 곳을 기준으로 하여 구분될 수 있고, 예측 로직의 개수가 복수인 경우에는 각각의 예측 로직을 기준으로 하여 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 결정할 수 있다. 따라서, 첫 번째 예측 로직을 기준으로 결정되는 적어도 하나의 제1 및 제2 레이어들과 두 번째 예측 로직을 기준으로 결정되는 적어도 하나의 제1 및 제2 레이어들은 서로 다를 수 있다. 기준이 되는 예측 로직이 다를 경우 적어도 하나의 제1 및 제2 레이어들에

따라 결정되는 적어도 하나의 제1 레이어의 가장 마지막 컨볼루션 레이어와 적어도 하나의 제2 레이어의 가장 최초 컨볼루션 레이어도 서로 다를 수 있다.

[0063] 컨볼루션 연산 중단 결정부(230)는 분류의 적정성을 결정하여 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있다. 분류의 적정성은 분류 결과를 실제로 사용하기에 적당한지를 수치화한 것에 해당할 수 있고 해당 모델을 통해 추론되는 결과에 대한 신뢰성에 대응될 수 있다. 컨볼루션 연산 중단 결정부(230)는 분류의 적정성이 특정 기준을 충족하는지를 검사하여 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있다.

[0064] 일 실시예에서, 컨볼루션 연산 중단 결정부(230)는 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 동적으로 결정하여 예측 로직의 연결성을 제어할 수 있다. 컨볼루션 연산 중단 결정부(230)는 컨볼루션 신경망 모델을 구성하는 컨볼루션 레이어들 사이에 연결되는 예측 로직이 복수인 경우 각각의 예측 로직을 기준으로 하여 적어도 하나의 제1 및 제2 컨볼루션 레이어들을 동적으로 결정할 수 있다. 또한, 컨볼루션 연산 중단 결정부(230)는 각 예측 로직에 의해 수행되는 예측 연산의 결과를 기초로 분류의 적정성을 결정하여 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있고, 특정 예측 로직에 대하여 컨볼루션 연산의 계속이 결정된 경우 다음 예측 로직까지의 컨볼루션 연산을 수행하여 예측 로직의 연결성을 제어할 수 있다. 즉, 예측 로직의 연결성은 컨볼루션 연산의 계속 결정으로 인하여 특정 예측 로직에서 다음 예측 로직까지의 컨볼루션 연산이 계속됨을 의미할 수 있다.

[0065] 일 실시예에서, 컨볼루션 연산 중단 결정부(230)는 정규화된 출력벡터의 제1 성분 및 제2 성분들 간의 차이를 적정 임계값과 비교함으로써 상기 분류의 적정성을 결정할 수 있다. 예를 들어, 컨볼루션 연산 중단 결정부(230)는 정규화된 출력벡터를 구성하는 성분들 중에서 가장 큰 값과 두 번째로 큰 값의 차이가 적정 임계값보다 큰 경우 예측 연산 결과를 신뢰하고 전체 모델의 수행을 중지할 수 있다. 따라서, 출력벡터의 구성 성분 중 가장 큰 값이 제1 성분에 해당하고 두 번째로 큰 값이 제2 성분에 해당할 수 있다. 또한, 적정 임계값은 분류의 적정성을 결정하는 기준 값에 해당할 수 있고, 입력 분류 장치(130)는 적정 임계값의 크기를 조정하여 컨볼루션 신경망의 사용 깊이를 가변적으로 조정할 수 있다.

[0066] 일 실시예에서, 컨볼루션 연산 중단 결정부(230)는 복수의 예측 로직들 각각에서 미리 예측된 분류의 적정성을 해당 컨볼루션 레이어에 의한 컨볼루션 연산과 동시에 수행하여 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있다. 컨볼루션 연산 중단 결정부(230)는 컨볼루션 신경망에 연결된 예측 로직들이 복수인 경우 각각의 예측 로직에서의 예측 연산을 해당 컨볼루션 레이어에 의한 컨볼루션 연산과 동시에 병렬적으로 수행할 수 있고, 예측 연산에 따른 분류의 적정성을 기초로 해당 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있다.

[0067] 일 실시예에서, 컨볼루션 연산 중단 결정부(230)는 컨볼루션 연산 순서에 따라 복수의 예측 로직들에서 미리 예측된 분류의 적정성을 차례대로 결정할 수 있다. 컨볼루션 연산 중단 결정부(230)는 컨볼루션 신경망을 구성하는 컨볼루션 연산 순서에 따라 복수의 예측 로직들에서의 예측 연산을 수행할 수 있고, 예측 연산에 따라 산출되는 분류의 적정성이 충분할 때까지 예측 로직에서의 예측 연산을 수행할 수 있다. 예를 들어, 컨볼루션 연산 중단 결정부(230)는 첫 번째 예측 로직에서 미리 예측된 분류의 적정성이 충분하지 않은 경우 두 번째 예측 로직에서 미리 예측된 분류의 적정성을 결정하여 컨볼루션 연산의 중단 또는 계속을 결정할 수 있다.

[0068] 일 실시예에서, 컨볼루션 연산 중단 결정부(230)는 예측 연산에 따른 분류의 적정성을 결정하여 가장 최초 컨볼루션 레이어와 직접 연결된 다음 컨볼루션 레이어를 통한 컨볼루션 연산의 지속적 수행 여부를 결정할 수 있다. 컨볼루션 연산 중단 결정부(230)는 예측 로직에서의 예측 연산에 따른 분류의 적정성을 결정할 수 있고, 분류의 적정성이 충분하지 않은 경우 컨볼루션 연산의 계속을 결정할 수 있다. 컨볼루션 연산 중단 결정부(230)에 의해 컨볼루션 연산의 계속이 결정된 경우 다음 컨볼루션 연산은 적어도 하나의 제2 컨볼루션 레이어들의 가장 최초 컨볼루션 레이어와 직접 연결된 다음 컨볼루션 레이어를 통한 컨볼루션 연산에 해당할 수 있다. 결과적으로, 예측 로직에 의한 예측 연산 이후 다음 컨볼루션 연산이 수행되는 곳은 적어도 하나의 제2 컨볼루션 레이어들의 가장 최초 컨볼루션 레이어가 아닌 해당 컨볼루션 레이어와 직접 연결된 다음 컨볼루션 레이어에 해당할 수 있다.

[0069] 제어부(250)는 입력 분류 장치(130)의 전체적인 동작을 제어하고, 컨볼루션 레이어부(210) 및 컨볼루션 연산 중단 결정부(230) 간의 제어 흐름 또는 데이터 흐름을 관리할 수 있다.

[0071] 도 3은 도 1에 있는 입력 분류 장치에서 수행되는 입력 분류 과정을 설명하는 순서도이다.

[0072] 도 3을 참조하면, 입력 분류 장치(130)는 컨볼루션 레이어부(210)를 통해 상호 연결되어 컨볼루션 연산을 수행하는 적어도 하나의 제1 및 제2 컨볼루션 레이어들 사이에 연결된 예측 로직에 의해 입력 데이터에 관한 분류를

미리 예측할 수 있다(단계 S310). 입력 분류 장치(130)는 컨볼루션 연산 중단 결정부(230)를 통해 분류의 적정성을 결정하여 컨볼루션 연산의 중단 또는 계속 여부를 결정할 수 있다(단계 S330).

[0074] 도 4는 도 1에 있는 입력 분류 장치에서 수행되는 추론 수행의 일 실시예를 설명하는 예시도이다.

[0075] 도 4를 참조하면, 입력 분류 장치(130)는 실행 중인 모델로부터 예측 지점의 임시 결과값을 추출해 분류기의 외부에서 동적으로 컨볼루션 신경망 분류기의 수행에 관여할 수 있다. 예측 지점(tapping point)은 전체 분류기 모델을 수행하기 전에 미리 결과를 예측할 수 있다는 의미에서 전결합층이 연결되는 지점에 해당할 수 있고 예측 로직이 연결되는 컨볼루션 레이어 간의 연결의 중간 지점에 해당할 수 있다. 입력 분류 장치(130)는 예측 로직에 의한 예측 연산을 통해 분류의 적정성을 결정할 수 있고, 분류의 적정성이 충분한 경우에는 실행 중인 모델을 중단하고 예측 로직에 의한 예측 연산의 결과를 최종 분류 결과로서 결정할 수 있다.

[0076] 입력 분류 장치(130)는 보조 분류기에 신뢰성 검토 알고리즘을 추가하여 해당 분류기에서 출력된 결과의 신뢰성을 판단하여 모델 수행 지속 여부를 결정할 수 있다. 분류 모델에 사용하는 소프트맥스는 다음의 수학식 1과 같이 계산될 수 있다.

[0078] [수학식 1]

[0079] $$Z_k = \frac{e^{a_k}}{\sum_{i=1}^{\#class}(e^{a_i})}$$

[0081] 여기에서, a는 전결합층에서 각 클래스에 대해 가중치를 연산한 결과로 길이가 클래스의 개수(#class)인 벡터에 해당할 수 있다. 즉, 소프트맥스는 분류의 전처리로서 전결합층에서 발생하는 결과값을 보정하여 각 클래스에 대한 함수의 결과값의 합이 1이 되도록 하는 함수에 해당할 수 있다. 입력값이 추론 과정에서 최종적으로 분류되는 클래스 k는 z_k = max(z)이며 이 때 반드시 a_k = max(a)이다. 따라서, 소프트맥스를 활성 함수로 적용하기 전의 전결합층에서 발생하는 결과값을 통해 미리 분류 결과를 확인할 수 있다. 전체 분류기 모델을 수행하기 전에 미리 결과를 예측할 수 있다는 의미에서 전결합층이 연결되는 지점은 예측 지점(tapping point)에 해당할 수 있다.

[0082] 예측 지점에서 결과가 발생했을 때, 추론의 신뢰성을 판단하기 위해 전결합층에서 발생하는 결과 벡터 a를 이용할 수 있다. 전결합층에서 발생하는 결과값은 입력 이미지에 따라 발생하는 결과값에 편차가 존재할 수 있다. 그러므로 발생하는 모든 결과값에 대해 동일한 연산을 적용하기 위해 신뢰성 판단을 수행하기 전에 얻어진 벡터를 다음의 수학식 2를 통해 정규화(normalize)할 수 있다.

[0084] [수학식 2]

[0085] $$y = \frac{a}{\sqrt{\sum_{i=1}^{\#class}(a_i{}^2)}}$$

[0087] 여기에서, k = argmax(y)라고 하면 벡터 a를 정규화한 벡터 y로 분류 결과를 획득할 수 있다. 즉, 예측 지점에서 연산을 수행하여 정규화한 벡터 y는 예측 벡터에 해당하고, k는 예측 분류 결과에 해당할 수 있다. 입력 분류 장치(130)는 임계값을 지정하여 연산한 차이값이 임계값보다 클 때 예측 분류 결과를 신뢰하고 전체 모델의 수행을 중지할 수 있다. 즉, 입력 분류 장치(130)는 임계값의 크기를 조정하여 컨볼루션 신경망의 사용 깊이를 가변적으로 조정할 수 있다. 이때, 임계값이 증가할수록 예측 결과를 신뢰할 확률은 감소하므로 임계값을 조절하는 방식으로 예측 분류 결과의 신뢰도 또한 조절할 수 있다.

[0089] 도 5는 도 2에 있는 컨볼루션 레이어부에 포함된 예측 로직에서 수행되는 예측 연산의 의사 코드를 설명하는 도면이다.

[0090] 도 5를 참조하면, 예측 로직(Prediction Logic)은 예측 지점에 연결된 보조 분류기의 전연결망으로부터 가중치를 추출하여 일차원으로 저장한 벡터 weights을 사용하여, 수행 중인 모델을 중단시킬지의 여부를 판단하면서 예측 지점에서 분류한 결과를 반환할 수 있다. 예측 로직은 이 벡터를 중간 출력값 flattened output과 행렬곱하여 벡터 a를 얻을 수 있다. 예측 로직은 얻어진 벡터 a를 정규화하여 예측 벡터 y를 생성할 수 있다. 예측 로직은 연산한 벡터 y의 가장 큰 값과 그 다음으로 큰 값의 차이를 임계값과 비교할 수 있다.

[0091] 입력 분류 장치(130)는 임계값보다 구한 차가 클 경우 예측으로 발생한 결과를 신뢰하며, 이 경우 남은 모델은

수행할 필요가 없다고 판단하여 수행되고 있는 모델 프로세스를 종료시키고 이 지점에서 예측한 결과를 반환할 수 있다. 입력 분류 장치(130)는 그렇지 않을 경우에는 예측한 결과를 버리고 다른 예측 지점에서 신뢰할 만한 결과가 발생하거나, 수행하고 있는 모델 프로세스가 종료되어 모델 결과가 발생하도록 할 수 있다. 입력 분류 장치(130)는 예측 로직에서 수행하는 알고리즘의 임계값을 조정하여 모델의 깊이를 가변적으로 조정할 수 있다.

[0093] 상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

[0095] 100: 컨볼루션 신경망 기반의 입력 분류 시스템

110: 사용자 단말 130: 입력 분류 장치

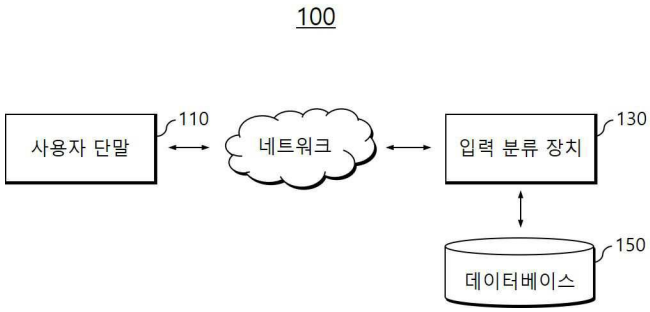
150: 데이터베이스

210: 컨볼루션 레이어부 230: 컨볼루션 연산 중단 결정부

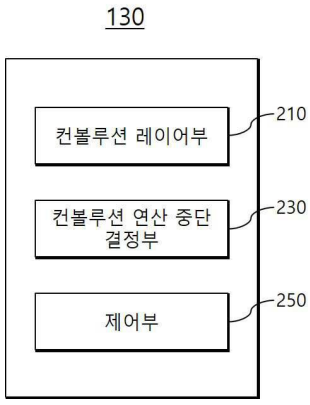
250: 제어부

도면

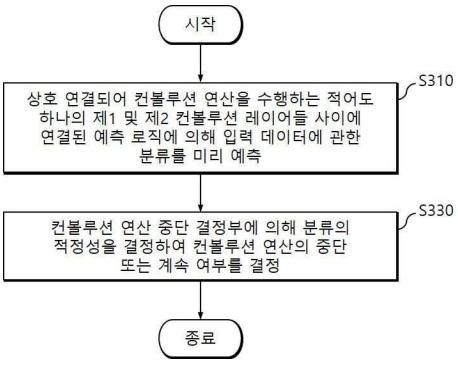
도면1



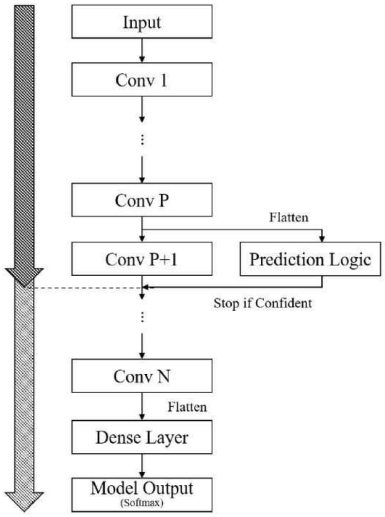
도면2



도면3



도면4



도면5

Input: W: m by n weight matrix, v : column vector of n elements (m is number of classes)
Output: Determine early termination and return classification label
Variable: a, y : column vector of m elements

$a = Wv$
$y = \text{normalize}(a)$ // $y = a/\sqrt{\sum_{i=1}^{\#class}(a_i^2)}$
$(top1, top2) = \text{choose\_top2}(y)$ // return two largest values
if $(top1 - top2 \geq threshold)$ **then**
    Terminate executing model
    Return ***k*** where $k = argmax(y)$
else
    Return ***none***