(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년12월04일
(11) 등록번호 10-2609945
(24) 등록일자 2023년11월30일

(51) 국제특허분류(Int. Cl.)
G06F 21/55 (2013.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 21/55 (2013.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2021-0115018
(22) 출원일자 2021년08월30일
심사청구일자 2021년08월30일
(65) 공개번호 10-2023-0032319
(43) 공개일자 2023년03월07일
(56) 선행기술조사문헌
Shiqing Ma et al., "NIC: Detecting Adversarial Samples with Neural Network Invariant Checking"(2019.02.)*
SuperDataScience, "Convolutional Neural Networks (CNN): Step 3 - Flattening" (2018.08.)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
윤지원
서울특별시 용산구 한강대로43길 8, 102동 1401호
최형준
부산광역시 동래구 충렬대로202번길 70, 226호 (수안동, 성림@)
(74) 대리인
김홍석

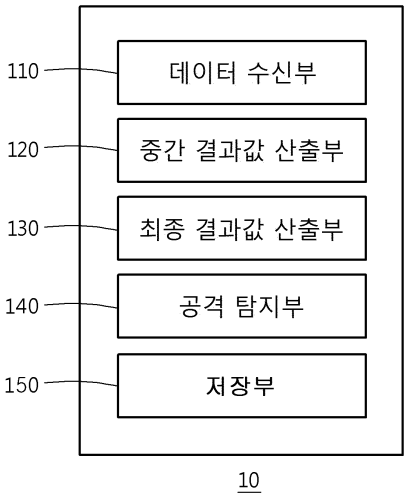
전체 청구항 수 : 총 3 항 심사관 : 정성훈

(54) 발명의 명칭 **딥러닝 기반의 적대적 공격 탐지 장치 및 방법**

(57) 요 약

적대적 공격 탐지 방법이 개시된다. 상기 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되고, 입력 데이터를 수신하는 단계, 상기 입력 데이터에 대한 탐지 모델의 출력값을 도출하는 단계, 및 상기 출력값에 기초하여 적대적 공격(adversarial attack)의 발생 여부를 판단하는 단계를 포함하고, 상기 탐지 모델은, 입력 레이어, 복수의 중간 레이어들, 및 출력 레이어를 포함하여 미리 정해진 동작을 수행하는 기본 모델과 상기 복수의 중간 레이어들 중 어느 하나의 중간 레이어의 출력에 대응하는 결과값을 출력하는 적어도 하나의 중간 출력 레이어를 포함하고, 상기 판단하는 단계는, 상기 출력 레이어의 출력값과 상기 적어도 하나의 중간 출력 레이어의 출력값의 동일성 여부에 기초하여 상기 적대적 공격의 발생 여부를 판단한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/082 (2023.01)

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 적대적 공격 탐지 방법에 있어서,

입력 데이터를 수신하는 단계;

상기 입력 데이터에 대한 탐지 모델의 출력값을 도출하는 단계; 및

상기 출력값에 기초하여 적대적 공격(adversarial attack)의 발생 여부를 판단하는 단계를 포함하고,

상기 탐지 모델은, 입력 레이어, 복수의 중간 레이어들, 및 제1 출력 레이어를 포함하여 미리 정해진 동작을 수행하는 기본 모델과 상기 복수의 중간 레이어들 중 어느 하나의 중간 레이어의 출력을 수신하여 결과값을 출력하는 적어도 하나의 제2 출력 레이어를 포함하고,

상기 판단하는 단계는, 상기 제1 출력 레이어의 출력값과 상기 적어도 하나의 제2 출력 레이어의 출력값의 동일성 여부에 기초하여 상기 적대적 공격의 발생 여부를 판단하고,

상기 판단하는 단계는, 상기 제1 출력 레이어의 출력값과 상기 적어도 하나의 제2 출력 레이어의 출력값이 동일하지 않은 경우 상기 적대적 공격이 발생한 것으로 판단하고,

상기 적어도 하나의 제2 출력 레이어는 m(m은 2이상의 자연수) 개이고,

상기 판단하는 단계는, 상기 m 개의 제2 출력 레이어들의 출력값들 중 n(n은 m보다 작거나 같은 자연수) 개의 출력값이 상기 제1 출력 레이어의 출력값과 동일하지 않은 경우 상기 적대적 공격이 발생한 것으로 판단하는,

적대적 공격 탐지 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 기본 모델은 MNIST(Modified National Institute of Standards and Technology) 모델 또는 CIFAR-10(Canadian Institute For advanced Research-10) 모델인,

적대적 공격 탐지 방법.

청구항 5

제1항에 있어서,

상기 적어도 하나의 제2 출력 레이어는 상기 어느 하나의 중간 레이어의 출력값을 1차원 벡터로 변환하는 플랫(Flat) 레이어를 포함하는,

적대적 공격 탐지 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 적대적 공격 탐지 방법에 관한 것으로, 특히 딥러닝 모델 또는 기계학습 모델을 대상으로 하는 적대

적 공격을 탐지하는 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 딥러닝 모델 또는 기계학습 모델에 예기치 않은 데이터를 입력하여 시스템이 의도하지 않는 결과를 출력하도록 하는 적대적 공격(adversarial attack)에 대한 대응책은 크게 탐지 기법과 방어 기법이 있다. 기존의 탐지 기법의 경우, 내부 구조나 알고리즘을 달리한 공격 탐지 모델을 별도로 구비하고 딥러닝 모델의 결과값과 공격 탐지 모델의 결과값을 비교하여 적대적 공격을 방어할 수 있다고 알려져 있다.

[0003] 그러나, 기존의 탐지 기법의 경우, 모델을 최초 학습할 때 기본 모델과 공격 탐지 모델 각각에 대하여 학습(즉, 두 번의 학습)하여야 하고, 내부 구조를 달리하거나 알고리즘을 달리하는 등의 노력이 필요하다. 또한, 모델을 업데이트할 경우에 또 다시 두 개의 모델을 별도로 업데이트 하여야 하며, 공격 탐지 모델을 보관하는 공간이 필요하다. 따라서, 하나의 모델에서 기본 모델이 제공하는 기능과 공격 탐지 기능을 모두 제공할 수 있는 방법을 제안하고자 한다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 대한민국 공개특허 제2018-0041953호 (2018.04.25. 공개)

(특허문헌 0002) 대한민국 공개특허 제2020-0095219호 (2020.08.10. 공개)

비특허문헌

[0005] (비특허문헌 0001) Weilin Xu, David Evans, and Yanjun Qi, "Feature squeezing: Detecting adversarial examples in deep neural network.", arXiv preprint arXiv:1704.01155 (2017)

발명의 내용

해결하려는 과제

[0006] 본 발명이 이루고자 하는 기술적인 과제는 딥러닝 모델을 대상으로 하는 적대적 공격을 탐지하는 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0007] 본 발명의 일 실시예에 따른 적대적 공격 탐지 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 적대적 공격 탐지 방법으로서, 입력 데이터를 수신하는 단계, 상기 입력 데이터에 대한 탐지 모델의 출력값을 도출하는 단계, 및 상기 출력값에 기초하여 적대적 공격(adversarial attack)의 발생 여부를 판단하는 단계를 포함하고, 상기 탐지 모델은, 입력 레이어, 복수의 중간 레이어들, 및 출력 레이어를 포함하여 미리 정해진 동작을 수행하는 기본 모델과 상기 복수의 중간 레이어들 중 어느 하나의 중간 레이어의 출력에 대응하는 결과값을 출력하는 적어도 하나의 중간 출력 레이어를 포함하고, 상기 판단하는 단계는, 상기 출력 레이어의 출력값과 상기 적어도 하나의 중간 출력 레이어의 출력값의 동일성 여부에 기초하여 상기 적대적 공격의 발생 여부를 판단한다.

발명의 효과

[0008] 본 발명의 실시예에 따른 적대적 공격 탐지 장치 및 방법에 의할 경우, 기본 딥러닝 모델이 탐지 기능을 수행하도록 함으로써, 모델 구조를 달리하거나 알고리즘을 달리하는 등의 노력 없이 적대적 공격을 탐지할 수 있는 효과가 있다.

[0009] 또한, 모델 업데이트가 필요한 경우에는 기존의 기본 모델만 업데이트하여도 되며 공격 탐지 모델을 보관할 별도의 공간을 마련할 필요가 없다.

도면의 간단한 설명

[0010] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른 적대적 공격 탐지 장치의 기능 블럭도이다.

도 2는 도 1에 도시된 공격 탐지 장치에 의해 수행되는 적대적 공격 탐지 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명에 의한 탐지 모델의 일 실시예를 도시한다.

도 4는 본 발명에 의한 탐지 모델의 다른 실시예를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0011] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0012] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0013] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0014] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0015] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0016] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0017] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0018] 도 1은 본 발명의 일 실시예에 따른 적대적 공격 탐지 장치의 기능 블럭도이고, 도 2는 도 1에 도시된 공격 탐지 장치에 의해 수행되는 적대적 공격 탐지 방법을 설명하기 위한 흐름도이다.

[0019] 도 1을 참조하면, 적대적 공격 탐지 장치(10)는 데이터 수신부(110), 중간 결과값 산출부(120), 최종 결과값 산출부(130), 및 공격 탐지부(140)를 포함한다. 실시예에 따라, 적대적 공격 탐지 장치(10)는 저장부(150)를 더 포함할 수 있다.

[0020] 적대적 공격 탐지 장치(10)는 수신된 입력 데이터에 대한 탐지 모델의 결과값들(최종 결과값과 적어도 하나의 중간 결과값)을 도출하고, 도출된 결과값들에 기초하여 적대적 공격을 탐지할 수 있다.

[0021] 데이터 수신부(110)는 유무선 통신망을 통하여 입력 데이터를 수신하거나, 소정의 입력 인터페이스를 통해 사용자로부터 입력 데이터를 수신할 수 있다. 실시예에 따라, 입력 데이터는 적대적 공격 탐지 장치(10)의 저장부(150)에 미리 저장되어 있을 수도 있다. 이 경우, 적대적 공격 탐지 장치(10)는 데이터 수신부(110)를 포함할지 않을 수도 있다.

[0022] 중간 결과값 산출부(120)는 입력 데이터에 대한 탐지 모델의 적어도 하나의 중간 결과값을 산출할 수 있고, 최종 결과값 산출부(130)는 입력 데이터에 대한 탐지 모델의 최종 결과값을 산출할 수 있다.

[0023] 적어도 하나의 중간 결과값과 최종 결과값은 입력 데이터에 대한 탐지 모델의 출력값이므로, 중간 결과값 산출부(120)와 최종 결과값 산출부(130)는 하나의 구성으로 이해될 수도 있다.

[0024] 탐지 모델은 사전 학습되어 미리 정해진 동작을 수행하는 기본 모델과 기본 모델에 포함된 중간 레이어의 출력에 기초하여 중간 결과값을 출력하는 적어도 하나의 중간 출력 레이어로 구성될 수 있다. 즉, 탐지 모델은 기본 모델에 포함된 복수의 중간 레이어들 중 어느 하나의 중간 레이어로부터 분기된 중간 출력층을 포함하는 것으로 이해될 수 있다. 중간 출력층은 복수개가 존재할 수 있다.

[0025] 기본 모델은 복수의 레이어들(입력 레이어, 복수의 중간 레이어, 및 출력 레이어)를 포함하는 딥러닝 모델 또는 기계 학습 모델을 의미할 수 있다. 실시예에 따라, 중간 레이어는 은닉 레이어(hidden layer)로 불릴 수도 있다. 기본 모델은 인공 신경망(Atrificial Neural Network, ANNN), 심층 신경망(Deep Neural Network, DNN), 콘볼루션 신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN), 심층 순환 신경망(deep RNN), RBM(Restricted Boltzmann Machine), DBN(Deep Belief Network), 생성 대립 신경망(Generative Adversarial Network, GAN), 장단기 메모리(Long Short Term Memory, LSTM), DCGAN(Deep Convolution GAN), GRU(Gated Recurrent Unit), AE(Auto Encoder), VAE(Variational AE), DAE(Denoising AE), SAE(Sparse AE), DCN(Deep Convolutional Network), DN(Deconvolutional Network), DCIGN(Deep Convolutional Inverse Graphics Network), LSM(Liquid State Machine), ELM(Extreme Learning Machine), ESN(Echo State Network), DRN(Deep Residual Network), CN(Capsule Network), DFN(Deep Feedforward Network) 등으로, 모델의 목적에 맞도록 학습 데이터를 이용하여 학습된 모델을 의미할 수 있다.

[0026] 입력 레이어는 입력 데이터를 복수의 중간 레이어들 중 첫번째 중간 레이어에 전송하는 역할을 수행한다. 출력 레이어는 복수의 중간 레이어들 중 마지막 중간 레이어의 출력을 입력받아 최종 결과값을 출력한다. 실시예에 따라, 출력 레이어와 마지막 중간 레이어 사이에는 플랫(Flat) 레이어를 더 포함할 수도 있다. 다른 실시예로, 플랫 레이어는 출력 레이어에 포함되는 구성으로 이해될 수도 있다. 플랫 레이어는 Flatten 함수를 이용하여 다차원(예컨대, 3차원) 텐서를 1차원 벡터로 변환할 수 있다.

[0027] 중간 출력 레이어는 기본 모델에 포함된 복수의 중간 레이어들 중 어느 하나의 중간 레이어의 출력을 수신하고 중간 결과값을 출력할 수 있다. 중간 출력 레이어는 복수개가 존재할 수 있다. 또한, 중간 출력 레이어와 중간 레이어 사이에는 플랫 레이어가 더 구비될 수 있다. 다른 실시예로, 플랫 레이어는 중간 출력 레이어에 포함되는 구성으로 이해될 수도 있다.

[0028] 공격 탐지부(140)는 적어도 하나의 중간 결과값과 최종 결과값의 비교를 통해 적대적 공격 발생 여부를 판단할 수 있다.

[0029] 중간 결과값이 하나인 경우, 공격 탐지부(140)는 중간 결과값과 최종 결과값이 동일한 경우 적대적 공격이 발생하지 않은 것으로 판단하고, 중간 결과값과 최종 결과값이 동일하지 않은 경우 적대적 공격이 발생했다고 판단할 수 있다.

[0030] 중간 결과값이 복수개인 경우, 공격 탐지부(140)는 복수의 중간 결과값들 중 적어도 하나와 최종 결과값이 동일한 경우 적대적 공격이 발생하지 않은 것으로 판단할 수 있다. 다시 말하면, 공격 탐지부(140)는 복수의 중간 결과값들 중 적어도 하나와 최종 결과값이 동일하지 않은 경우, 적대적 공격이 발생한 것으로 판단할 수 있다.

[0031] 구체적인 예로, 중간 결과값의 개수가 m(m은 2 이상의 자연수)개일 때, 공격 탐지부(140)는 n(n은 m보다 작거나 같은 자연수)보다 크거나 같은 개수의 중간 결과값과 최종 결과값이 동일하지 않은 경우 적대적 공격이 발생한 것으로 판단할 수 있다. m이 5이고 n이 2인 경우, 공격 탐지부(140)는 5개의 중간 결과값들 중 2보다 크거나 같은 개수(2개, 3개, 4개, 또는 5개)의 중간 결과값들이 최종 결과값과 동일하지 않은 경우 적대적 공격이 발생한

것으로 판단할 수 있다. 이와는 다르게, 공격 탐지부(140)는 2보다 작은 개수(1개 또는 0개)의 중간 결과값이 최종 결과값과 동일하지 않은 경우, 적대적 공격이 발생하지 않은 것으로 판단할 수 있다.

[0032] 여기서, m과 n의 구체적인 값은 사용 환경에 따라 유동적으로 조절이 가능하다. 요구되는 보안의 정도가 작다면, m의 값을 작게 설정하거나, 및/또는 n의 값을 크게 설정할 수 있다. 이와는 다르게, 요구되는 보안의 정도가 크다면, m의 값을 크게 설정하거나, 및/또는 n의 작게 설정할 수 있다.

[0033] 저장부(150)에는 입력 데이터, 탐지 모델(기본 모델을 포함할 수 있음), 적어도 하나의 중간 결과값, 최종 결과값, 공격 탐지의 결과, 공격 탐지 과정에서 일시적으로 또는 비일시적으로 생성되는 데이터 등이 저장될 수 있다.

[0034] 적대적 공격 탐지 장치(10)는 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치로 구현될 수 있다. 따라서, 적대적 공격 탐지 방법을 구성하는 각 단계의 적어도 일부는 컴퓨팅 장치에 구비된 프로세서의 동작으로 인식될 수도 있다.

[0035] 도 1에 도시된 적대적 공격 탐지 장치(10)의 구성들 각각은 기능 및 논리적으로 분리될 수 있음으로 나타내는 것이며, 반드시 각각의 구성이 별도의 물리적 장치로 구분되거나 별도의 코드로 작성됨을 의미하는 것이 아님을 본 발명의 기술분야의 평균적 전문가는 용이하게 추론할 수 있을 것이다.

[0036] 또한, 본 명세서에서 "~부"라 함은, 본 발명의 기술적 사상을 수행하기 위한 하드웨어 및 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적, 구조적 결합을 의미할 수 있다. 예컨대, 상기 모듈은 소정의 코드와 상기 소정의 코드가 수행되기 위한 하드웨어 리소스의 논리적인 단위를 의미할 수 있으며, 반드시 물리적으로 연결된 코드를 의미하거나, 한 종류의 하드웨어를 의미하는 것이 아니다.

[0037] 도 3은 본 발명에 의한 탐지 모델의 일 실시예를 도시하고, 도 4는 본 발명에 의한 탐지 모델의 다른 실시예를 도시한다. 구체적으로, 도 3에 도시된 탐지 모델에 포함된 기본 모델은 MNIST(Modified National Institute of Standards and Technology) 모델이고, 도 4에 도시된 탐지 모델에 포함된 기본 모델은 CIFAR-10(Canadian Institute For Advanced Research-10) 모델이다.

[0038] 도 3을 참조하면, 탐지 모델은 입력 레이어, 두 개의 중간 레이어, 및 출력 레이어를 포함하는 기본 모델(MNIST 모델)과 제1 중간 레이어로부터 분기된 하나의 중간 출력 레이어를 포함한다. 여기서, 출력 레이어의 출력값(최종 결과값)과 중간 출력 레이어의 출력값(중간 결과값)이 동일한 경우 적대적 공격이 발생하지 않은 것으로 판단되고, 최종 결과값과 중간 결과값이 동일하지 않은 경우 적대적 공격이 발생한 것으로 판단될 수 있다.

[0039] 도 4를 참조하면, 탐지모델은 입력 레이어, 9개의 중간 레이어, 및 출력 레이어를 포함하는 기본 모델(CIFAR-10)과 제3 중간 레이어로부터 분기된 제1 중간 출력 레이어와 제6 중간 레이어로부터 분기된 제2 중간 출력 레이어를 포함한다.

[0040] 도 4의 예에서, 공격 탐지부(140)는 m(m=2)개의 중간 결과값 중 n보다 크거나 같은 개수의 중간 결과값이 최종 결과값과 동일하지 않은 경우를 적대적 공격이 발생한 경우로 판단할 수 있다. 여기서, n은 m보다 작거나 같은 자연수이므로, 1 또는 2의 값을 가질 수 있다.

[0041] n이 1일 때, 1개의 중간 결과값 또는 2개의 중간 결과값이 최종 결과값과 동일하지 않은 경우 적대적 공격이 발생한 것으로 판단될 수 있고, 0개의 중간 결과값이 최종 결과값과 동일하지 않은 경우(즉, 모든 중간 결과값이 최종 결과값과 동일한 경우) 적대적 공격이 발생하지 않은 것으로 판단될 수 있다.

[0042] n이 2일 때, 2개의 중간 결과값이 최종 결과값과 최종 결과값과 동일하지 않은 경우 적대적 공격이 발생한 것으로 판단될 수 있고, 0개의 중간 결과값 또는 1개의 중간 결과값이 최종 결과값과 동일하지 않은 경우 적대적 공격이 발생하지 않은 것으로 판단될 수 있다.

[0043] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사

용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0044] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0045] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0046] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0047] 10 : 적대적 공격 탐지 장치

110 : 데이터 수신부

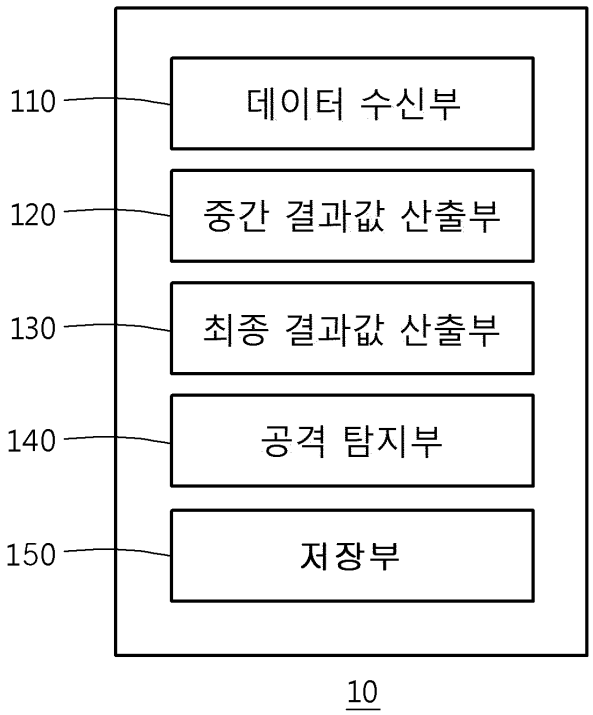
120 : 중간 결과값 산출부

130 : 최종 결과값 산출부

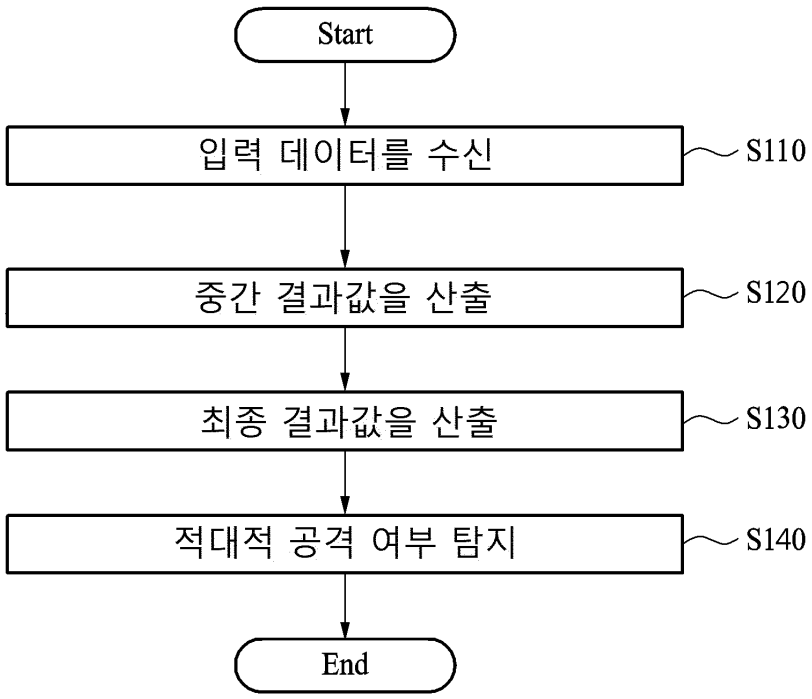
150 : 저장부

도면

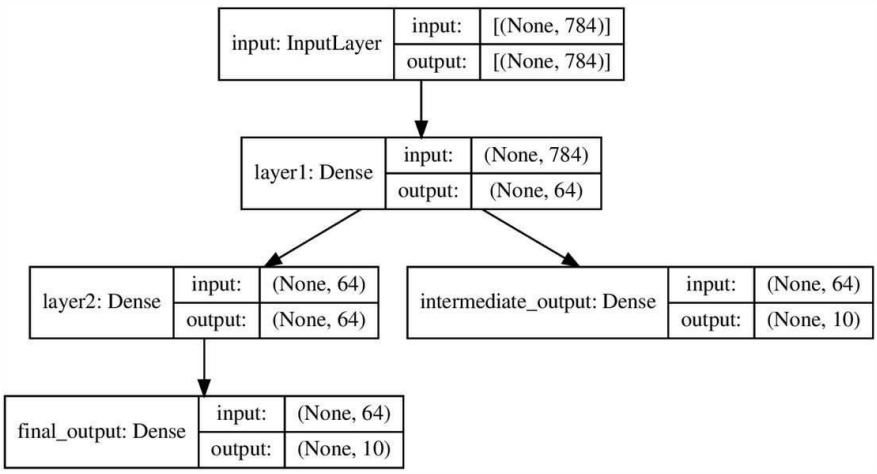
도면1


110
데이터 수신부
120
중간 결과값 산출부
130
최종 결과값 산출부
140
공격 탐지부
150
저장부
10

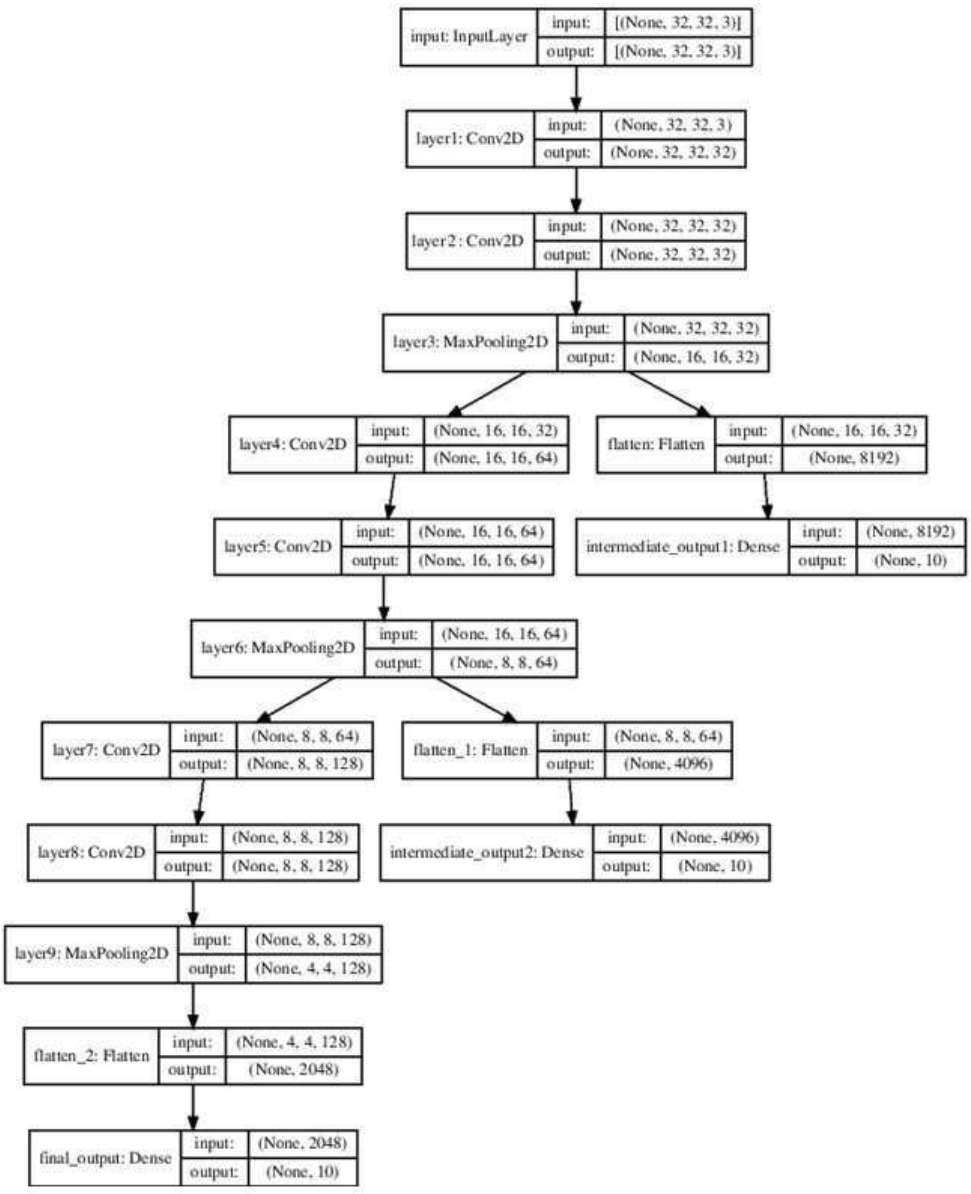
도면2


Start
입력 데이터를 수신
S110
중간 결과값을 산출
S120
최종 결과값을 산출
S130
적대적 공격 여부 탐지
S140
End

도면3


input: InputLayer
input: [(None, 784)]
output: [(None, 784)]
layer1: Dense
input: (None, 784)
output: (None, 64)
layer2: Dense
input: (None, 64)
output: (None, 64)
intermediate_output: Dense
input: (None, 64)
output: (None, 10)
final_output: Dense
input: (None, 64)
output: (None, 10)

도면4


input: InputLayer
input: [(None, 32, 32, 3)]
output: [(None, 32, 32, 3)]
layer1: Conv2D
input: (None, 32, 32, 3)
output: (None, 32, 32, 32)
layer2: Conv2D
input: (None, 32, 32, 32)
output: (None, 32, 32, 32)
layer3: MaxPooling2D
input: (None, 32, 32, 32)
output: (None, 16, 16, 32)
layer4: Conv2D
input: (None, 16, 16, 32)
output: (None, 16, 16, 64)
flatten: Flatten
input: (None, 16, 16, 32)
output: (None, 8192)
layer5: Conv2D
input: (None, 16, 16, 64)
output: (None, 16, 16, 64)
intermediate_output1: Dense
input: (None, 8192)
output: (None, 10)
layer6: MaxPooling2D
input: (None, 16, 16, 64)
output: (None, 8, 8, 64)
layer7: Conv2D
input: (None, 8, 8, 64)
output: (None, 8, 8, 128)
flatten_1: Flatten
input: (None, 8, 8, 64)
output: (None, 4096)
layer8: Conv2D
input: (None, 8, 8, 128)
output: (None, 8, 8, 128)
intermediate_output2: Dense
input: (None, 4096)
output: (None, 10)
layer9: MaxPooling2D
input: (None, 8, 8, 128)
output: (None, 4, 4, 128)
flatten_2: Flatten
input: (None, 4, 4, 128)
output: (None, 2048)
final_output: Dense
input: (None, 2048)
output: (None, 10)