(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0013654
(43) 공개일자 2024년01월30일

(51) 국제특허분류(Int. Cl.)
G06N 3/09 (2023.01) *G06N 5/04* (2023.01)
G06V 10/764 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06N 3/09 (2023.01)
G06N 5/045 (2023.01)
(21) 출원번호 10-2023-0083509
(22) 출원일자 2023년06월28일
심사청구일자 2023년06월28일
(30) 우선권주장
1020220091273 2022년07월22일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이상근
경기도 하남시 덕풍서로 65, 503동 103호
이정현
서울특별시 성북구 안암로 145, 고려대학교안암캠퍼스 로봇융합관212호
한성민
서울특별시 동대문구 무학로45길 10, 401호
(74) 대리인
김등용

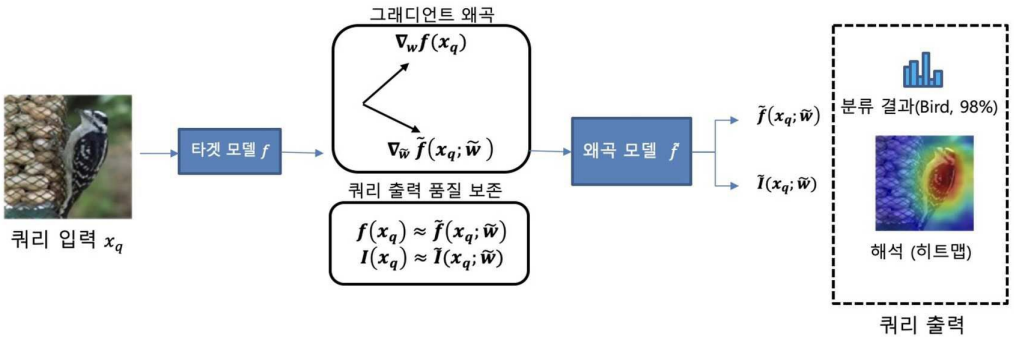
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 **딥러닝 기반 분류 시스템에 대한 모델 탈취 방어 장치및 방법**

(57) 요 약

딥러닝 기반 분류 시스템에 대한 모델 탈취 방어 장치 및 방법이 개시된다. 상기 모델 탈취 방어 방법은 적어도 프로세서(processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 쿼리 입력을 수신하는 단계, 상기 쿼리 입력에 대한 분류 동작을 수행하는 분류 모델에 대한 그래디언트(gradient)를 왜곡하여 왜곡 모델을 생성하는 단계, 및 상기 왜곡 모델의, 상기 쿼리 입력에 대한 쿼리 출력을 생성하는 단계를 포함한다.

대 표 도 - 도5



(52) CPC특허분류
G06V 10/764 (2023.08)
G06V 10/82 (2022.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711160323
과제번호	2020-0-01749-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	스마트미디어 서비스 혁신과 발전을 위한 융합연구
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

적어도 프로세서(processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 모델 탈취 방어 방법에 있어서,

쿼리 입력(x)을 수신하는 단계;

상기 쿼리 입력에 대한 분류 동작을 수행하는 분류 모델에 대한 그래디언트(gradient)를 왜곡하여 왜곡 모델을 생성하는 단계; 및

상기 왜곡 모델의, 상기 쿼리 입력에 대한 쿼리 출력을 생성하는 단계를 포함하는 모델 탈취 방어 방법.

청구항 2

제1항에 있어서,

상기 왜곡 모델을 생성하는 단계는 상기 분류 모델에 포함된 레이어(layer)들 중 적어도 하나의 레이어의 그래디언트를 왜곡하는,

모델 탈취 방어 방법.

청구항 3

제1항에 있어서,

상기 분류 모델은 분류 결과와 상기 분류 결과에 대한 해석(interpretation)을 함께 제공하는 분류 모델인,

모델 탈취 방어 방법.

청구항 4

제1항에 있어서,

상기 왜곡 모델을 생성하는 단계는,

수학식 1로 정의되는 상기 왜곡 모델의 손실 함수($L_{orth}(\cdot)$)를 최소화하기 위한 목적 함수(objective function)를 이용한 최적화 문제(optimization problem)를 풀이하여 상기 왜곡 함수를 생성하고,

상기 수학식 1은 $L_{\text{orth}}(\tilde{w}) = \cos\angle(\nabla_{\tilde{w}}\tilde{f}(x;\tilde{w}), \nabla_{w}f(x;w))$ 이고,

상기 $f(x;w)$는 상기 분류 모델의 분류 결과를 나타내고,

상기 $\tilde{f}(x;\tilde{w})$는 상기 왜곡 모델의 분류 결과를 나타내는,

모델 탈취 방어 방법.

청구항 5

제4항에 있어서,

상기 왜곡 모델을 생성하는 단계는,

수학식 2로 정의되는 제1 제약 조건을 이용하여 상기 왜곡 함수를 생성하고,

상기 수학식 2는 $\tilde{f}(x;\tilde{w})_{s_1} \geq \cdots \geq \tilde{f}(x;\tilde{w})_{s_k} \geq \max_{j \in S'} f(x;\tilde{w})_j$ 이고,

상기 s_i(i는 1 보다 크거나 같고 K 보다 작거나 같은 자연수)는 상기 분류 모델의 i 번째로 큰 확률 값을 갖는

클래스 인덱스를 나타내고,

상기 S′ 은 k(k는 1 보다 크거나 같고 K 보다 작거나 같은 자연수, K는 임의의 자연수) 번째로 확률값이 큰 클래스까지를 제외한 나머지 클래스의 집합을 나타내는,

모델 탈취 방어 방법.

청구항 6

제5항에 있어서,

상기 왜곡 모델을 생성하는 단계는,

수학식 3으로 정의되는 제2 제약 조건을 이용하여 상기 왜곡 함수를 생성하고,

상기 수학식 3은 $\tilde{I}(x;\tilde{w})_{a_1} \geq \tilde{I}(x;\tilde{w})_{a_2} \geq \cdots \geq \tilde{I}(x;\tilde{w})_{a_n}$ 이고,

상기 $\tilde{I}(x;\tilde{w})$는 상기 쿼리 입력에 대한 상기 왜곡 모델의 분류에 대한 해석이고,

상기 a_i(i는 1 보다 크거나 같고 n 보다 작거나 같은 자연수, n은 임의의 자연수)는 상기 분류 모델의 해석($I(x;w)$)에서 i 번째로 큰 원소 값의 인덱스인,

모델 탈취 방어 방법.

청구항 7

제6항에 있어서,

상기 왜곡 모델을 생성하는 단계는,

상기 제1 제약 조건과 상기 제2 제약 조건을 포함하는 제약 조건을 갖는 최적화 문제(Constrained Optimization Problem, COP)를 벌칙 기법(penalty method)을 적용하여 제약 조건을 갖지 않는 최적화 문제로 변환한 후에 상기 왜곡 모델을 생성하는,

모델 탈취 방어 방법.

청구항 8

제1항에 있어서,

상기 왜곡 모델을 생성하는 단계는,

상기 왜곡 모델에 포함된 레이어들 중 민감도가 가장 큰 적어도 하나의 레이어를 선택하여 그래디언트를 왜곡시키는,

모델 탈취 방어 방법.

청구항 9

제8항에 있어서,

상기 왜곡 모델을 생성하는 단계는,

상기 왜곡 모델에 포함된 레이어들 중 누적 민감도가 m(m은 0 보다 크거나 같고 100 보다 작거나 같은 임의의 실수)%가 되는 최소 개수의 레이어를 그래디언트 왜곡에 참여시키는,

모델 탈취 방어 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 딥러닝 시스템을 대상으로 하는 모델 탈취 공격(model stealing attack)에 대한 방어 기법에 관한 것

이다.

배 경 기 술

[0002] 본 발명은 블랙박스(blackbox) 형태로 분류 문제를 해결하는 딥러닝 기반 분류 시스템 또는 딥러닝 기반 분류 모델을 대상으로 한다. 여기서, 블랙박스 형태는 분류 모델의 구조, 학습 파라미터 등이 외부에게 감추어진 형태를 의미한다. 이때, 도 1에 도시된 바와 같이, 사용자는 쿼리 입력(query input)을 전송하고 해당 입력에 대한 분류 결과를 쿼리 출력(query output)으로 수신함으로써 분류 모델을 사용할 수 있다.

[0003] 딥러닝 기반 분류 모델의 경우 학습에 매우 많은 비용(ex: 대규모의 라벨링된 데이터셋)이 소모될 뿐 아니라 모델 자체가 부가가치가 높은 지적 재산으로써의 역할을 하기 때문에 대부분의 딥러닝 기반 분류 시스템이 블랙박스 형태로 서비스 된다.

[0004] 블랙박스 형태로 서비스되는 딥러닝 기반 분류 시스템 내부의 분류 모델의 기능을 탈취하려는 공격을 모델 탈취 공격(model stealing attack)이라고 칭한다. 또한, 공격자의 탈취 목표가 되는 딥러닝 기반 분류 시스템 내의 분류 모델을 타겟 모델이라고 칭한다. 모델 탈취 공격의 대략적인 과정은 다음과 같다.

[0005] 1. 모델 탈취 공격자가 라벨링 되지 않은 데이터셋을 확보 (공공 데이터셋 등)

[0006] 2. 모델 탈취 공격자가 임의의 공격 모델을 정의 (구조가 반드시 타겟 모델과 일치할 필요 없음)

[0007] 3. 모델 탈취 공격자가 공격자 데이터셋 내의 데이터들을 쿼리 입력으로써 타겟 모델(탈취 목표)에 전송

[0008] 4. 전송한 쿼리 입력에 대한 쿼리 출력을 수신함

[0009] 5. 수신된 쿼리 출력을 활용해 데이터셋을 라벨링하고 라벨링된 데이터셋을 이용하여 공격 모델 학습

[0010] 6. 공격 모델의 분류 성능이 타겟 모델과 비슷해질수록 모델 탈취 공격이 성공적으로 이루어졌다고 평가할 수 있음

[0011] 최근, 도 2에 도시된 바와 같이, 쿼리 출력으로 분류 결과 뿐 아니라 해당 분류 결과에 대한 해석(interpretation)까지 제공해 주는 딥러닝 기반 분류 시스템이 활발히 도입되고 있다. 모델 탈취 공격 과정에서 분류 결과 뿐 아니라 해석까지 악용할 경우 더 높은 탈취 성능을 달성할 수 있다는 연구가 존재한다. 이때, 해석을 악용한 모델 탈취 공격 과정은 전술한 과정과 동일하나 공격 모델 학습 과정(5.)에서 해석을 비슷하게 만드는 메커니즘을 추가하게 된다.

[0012] 모델 탈취 방어를 위하여 타겟 모델의 분류 결과에 노이즈를 첨가하는 방어 기술에 대한 연구들이 존재한다. 그러나, 기존 연구에 의한 방어 기술들은 모두 분류 결과만 악용하는 모델 탈취 공격에 집중 하고 있으며 해석을 악용한 모델 탈취 공격에 대한 방어 기술 및 연구는 현재 전무하다.

[0013] 이에, 본 발명에서는 딥러닝 기반 시스템을 위한 모델 탈취 방어 기법을 제안한다. 제안하는 모델 탈취 방어 기법은 그래디언트 왜곡(gradient misdirection)을 통해서 모델 탈취 공격을 저지하는 동시에 딥러닝 시스템의 분류 성능과 해석 품질을 높은 수준으로 유지할 수 있다. 또한, 타겟 딥러닝 기반 시스템은 해석을 동반하는 시스템과 해석을 동반하지 않는 시스템 모두를 포함할 수 있다.

선행기술문헌

특허문헌

[0014] (특허문헌 0001) 대한민국 공개특허 제2022-0025455호 (2022.03.03. 공개)

(특허문헌 0002) 대한민국 등록특허 제2170636호 (2020.10.27. 공고)

(특허문헌 0003) 대한민국 공개특허 제2022-0043023호 (2022.04.05. 공개)

발명의 내용

해결하려는 과제

[0015] 본 발명이 이루고자 하는 기술적인 과제는 딥러닝 기반 분류 시스템에 대한 모델 탈취 방어 장치 및 방법을 제

공하는 것이다.

과제의 해결 수단

[0016] 본 발명의 일 실시예에 따른, 모델 탈취 방어 방법은 적어도 프로세서(processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 모델 탈취 방어 방법으로써, 쿼리 입력(x)을 수신하는 단계, 상기 쿼리 입력에 대한 분류 동작을 수행하는 분류 모델에 대한 그래디언트(gradient)를 왜곡하여 왜곡 모델을 생성하는 단계, 및 상기 왜곡 모델의, 상기 쿼리 입력에 대한 쿼리 출력을 생성하는 단계를 포함한다.

발명의 효과

[0017] 본 발명의 실시예에 따른 딥러닝 기반 분류 시스템에 대한 모델 탈취 방어 장치 및 방법에 의할 경우, 분류 결과에 대한 신뢰성과 해석가능성을 모두 확보할 수 있다.

[0018] 또한, 그래디언트 왜곡(gradient misdirection)을 통해 타겟 모델의 분류 결과 및/또는 분류에 대한 해석을 동시에 보호할 수 있다.

도면의 간단한 설명

[0019] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 딥러닝 기반 분류 시스템의 동작을 설명하기 위한 도면이다.

도 2는 해석을 함께 제공하는 딥러닝 기반 분류 시스템의 동작을 설명하기 위한 도면이다.

도 3은 왜곡 모델의 top-3 분류 결과 보존 예시를 도시한다.

도 4는 왜곡 모델의 분류 결과 해석의 품질 보존 예시를 도시한다.

도 5는 본 발명의 일 실시예에 따른 모델 탈취 방어 방법의 개요도이다.

도 6은 본 발명의 일 실시예에 따른 모델 탈취 방어 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 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.

[0021] 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.

[0022] 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.

[0023] 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.

[0024] 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, 구

성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0025] 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.

[0026] 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.

[0027] 본 발명은 분류 결과와 분류 결과에 대한 해석(interpretation)을 함께 쿼리 출력으로 제공하는 딥러닝 기반 시스템 및/또는 분류 결과만을 쿼리 출력으로 제공하는 딥러닝 기반 시스템을 위한 모델 탈취 방어 기술에 관한 발명이다. 제안하는 모델 탈취 방어 기술은 그래디언트 왜곡(gradient misdirection)을 통해서 모델 탈취 공격을 저지하고, 및/또는 딥러닝 시스템의 분류 성능과 해석 품질을 높은 수준으로 유지할 수 있다.

[0028] 다수의 딥러닝 시스템들은 분류 결과에 대한 신뢰성과 해석가능성을 확보하기 위해 분류 모델을 XAI(eXplainable Artificial Intelligence) 기법과 결합하여 서비스 한다. 이때, 딥러닝 시스템의 쿼리 출력은 타겟 모델의 분류 결과와 XAI 기법으로 생성된 분류 결과에 대한 해석(interpretation)으로 구성된다. 최근, 분류 결과와 해석을 함께 모델 탈취 공격에 악용할 경우, 분류 결과만 악용했을 때 보다 타겟 모델을 더 효과적으로 복제할 수 있다는 연구 결과가 발표되었다. 하지만 종래의 모델 탈취 방어 기술들은 타겟 모델의 분류 결과만을 활용하는 모델 탈취 공격을 상정하고 있으며 분류 결과의 해석에 대한 보호는 제공하지 않는다. 본 발명은 타겟 모델에 대한 분류 결과와 해석 모두에 대한 모델 탈취 방어를 제공하는 최초의 모델 탈취 방어 기술이라 할 수 있다. 다만, 본 발명의 권리범위가 타겟 모델에 제한되는 것은 아니며, 해석을 동반하지 않는 분류 모델에 대한 모델 탈취 방어 기법을 제한하는 것은 아니다.

[0029] 본 발명은 그래디언트 왜곡(gradient misdirection)을 활용하여 타겟 모델의 분류 결과 및/또는 분류 결과에 대한 해석을 모델 탈취 공격으로부터 보호한다. 종래 모델 탈취 방어 기술들은 모델 탈취 공격자의 특정 정보(예컨대, 공격 모델의 그래디언트 정보, 공격자의 데이터셋 유형)를 알거나 유추할 수 있다는 가정에 의존하고 있다. 하지만, 이러한 방어 기술들의 경우 가정이 성립하지 않을 때 방어 성능이 하락할 가능성이 매우 높으며 이를 악용한 적응형 공격(adaptive attack)에 대해서도 취약하게 된다. 본 발명에서 활용하는 그래디언트 왜곡은 타겟 모델에 대한 정보만 활용하기 때문에 모델 탈취 공격자에 대한 특정 가정에 의존하지 않고도 안정적인 모델 탈취 방어 성능을 달성 할 수 있다.

[0030] 모델 탈취 방어 기술을 적용할 경우 필연적으로 타겟 모델의 쿼리 출력에 변형이 가해진다. 종래의 모델 탈취 방어는 타겟 모델의 top-1 분류 정확도 측면에서만 쿼리 출력의 품질을 유지하는 반면, 본 발명은 top-k(k는 1보다 크거나 같고 K 보다 작거나 같은 임의의 자연수, K는 분류 결과인 클래스의 수로써 임의의 자연수) 분류 정확도 및/또는 분류 결과에 대한 해석 품질을 보존하기 위한 새로운 메커니즘을 포함한다.

[0031] 이하에서는, 본 발명의 실시예들에 대해 보다 상세하게 설명한다.

[0032] **(문제 정의)**

[0033] 본 발명에서 다루는 분류 문제의 입력을 $x \in R^n$, 클래스의 수를 K 라고 가정한다.

[0034] 모델 탈취 공격 대상이 되는 타겟 모델의 함수를 $f(\cdot;w)$, 모델 탈취 공격자가 사용하는 공격 모델의 함수를 $f_A(\cdot;w_A)$로 표현한다. 이때, w와 w_A는 각각 타겟 모델과 공격 모델의 학습 파라미터를 의미한다. 주어진 입력 x에 대해서 모델 함수의 출력 $f(x;w)$, $f_A(x;w_A)$은 크기가 K인 소프트맥스 출력이며, 출력의 k 번째 원소는 k 번째 클래스에 대한 확률값을 나타낸다.

[0035] 타겟 모델과 공격 모델의 분류 결과에 대한 해석을 생성하는 해석기(interpreter) 함수를 각각 $I(\cdot;w)$와 $I_A(\cdot;w_A)$로 표현한다. 주어진 입력 x에 대해서 해석기의 출력 $I(x;w)$, $I_A(x;w_A)$은 입력과 동일한 크기의 실수값으로 구성된 히트맵이라고 가정한다.

[0036] 본 발명에서 분류 모델 서비스 제공자(모델 탈취 방어자)는 주어진 쿼리 입력 x에 대해 쿼리 출력($f(x;w)$, $I(x;w)$)을 제공하며 모델 탈취 공격자는 라벨링이 되지 않은 공격 데이터를 쿼리 입력으로 타겟 모델에 전송하여 쿼리 출력을 제공받은 뒤, 쿼리 출력을 통해 공격 데이터를 라벨링하여 공격 모델 $f_A(\cdot;w_A)$과 그에 대한 해석기 $I_A(\cdot;w_A)$를 학습한다.

[0037] **(그래디언트 왜곡)**

[0038] 본 발명에서는 주어진 쿼리 입력 x에 대한 본래 타겟 모델의 쿼리 출력($f(x;w)$, $I(x;w)$)을 직접 노출시키는 대신 왜곡 모델(misdirection model)의 쿼리 출력$(\tilde{f}(x;\tilde{w}),\tilde{I}(x;\tilde{w}))$을 노출 시킨다. 이때, 왜곡 모델은 모델 탈취 공격을 효과적으로 저지하기 위하여 모델 탈취 공격에 필요한 핵심적인 타겟 모델 정보인 학습 파라미터(예컨대, 가중치)에 대한 그래디언트 정보를 왜곡 시키도록 생성된다.

[0039] 왜곡 모델 $\tilde{f}(\cdot;\tilde{w})$과 그에 대한 해석기 $\tilde{I}(\cdot;\tilde{w})$의 학습 파라미터에 대한 그래디언트 정보를 왜곡시키기 위해서 왜곡 모델의 그래디언트 벡터의 $\nabla_{\tilde{w}}\tilde{f}(x;\tilde{w})$ 방향이 본래 타겟 모델의 그래디언트 벡터 $\nabla_w f(x;w)$의 방향과 직교하도록 유도한다. 두 벡터를 직교하게 하기 위한 왜곡 모델의 손실 함수를 다음 수학식 1과 같이 나타낼 수 있다.

[0040] [수학식 1]

[0041] $$L_{\text{orth}}(\tilde{w}) = \cos\angle(\nabla_{\tilde{w}}\tilde{f}(x;\tilde{w}), \nabla_w f(x;w))$$

[0042] **(쿼리 출력 품질 보존)**

[0043] 그래디언트 왜곡을 위한 손실 함수 L_{orth}만을 최소화하여 왜곡 모델을 생성할 경우 왜곡 모델의 쿼리 출력 $(\tilde{f}(x;\tilde{w}),\tilde{I}(x;\tilde{w}))$의 품질을 보장할 수 없다. 따라서, 본 발명에서는 왜곡 모델의 쿼리 출력의 품질(분류 정확도 및/또는 분류 결과의 해석가능성)을 보존하기 위해 분류 결과 $\tilde{f}(x;\tilde{w})$와 분류 결과에 대한 해석 $\tilde{I}(x;\tilde{w})$에 대한 제약조건을 추가한다.

[0044] Top-k 분류 정확도를 보존하기 위하여 주어진 쿼리 입력 x에 대해 왜곡 모델의 가장 큰 확률값을 갖는 클래스부터 k(k는 1 보다 크거나 같고 K 보다 작거나 같은 자연수, K는 클래스의 수) 번째로 큰 확률값을 갖는 클래스까지의 인덱스를 본래 타겟 모델과 동일하도록 한다. 본래 타겟 모델의 i 번째로 큰 확률 값의 클래스 인덱스를 s_i(i는 1 보다 크거나 같고 K 보다 작거나 같은 자연수), k 번째로 확률값이 큰 클래스 인덱스 까지를 제외한 나머지 클래스의 집합을 $S' := \{1,\cdots,K\}\setminus\{s_1,\cdots s_k\}$로 정의할 때 top-k 분류 정확도를 보존하기 위한 제약 조건을 다음 수학식 2와 같이 나타낼 수 있다.

[0045] [수학식 2]

[0046] $$\tilde{f}(x;\tilde{w})_{s_1} \geq \cdots \geq \tilde{f}(x;\tilde{w})_{s_k} \geq \max_{j\in S'} f(x;\tilde{w})_j$$

[0047] 분류 결과 해석의 품질을 보존하기 위하여 주어진 쿼리 입력 x에 대해 왜곡 모델의 분류에 대한 해석 $\tilde{I}(x;\tilde{w})$의 각 원소의 크기에 따른 순서를 본래 타겟 모델과 동일하도록 한다. 본래 타겟 모델의 해석 $I(x;w)$에서 i 번째로 큰 원소 값의 인덱스를 a_i(i는 1 보다 크거나 같고 n 보다 작거나 같은 자연수, n은 입력의 차원 또는 해석의 차원(크기)를 나타내는 자연수)라고 할 때, 왜곡 모델의 해석 품질 보존을 위한 제약 조건을 다음 수학식 3과 같이 나타낼 수 있다.

[0048] [수학식 3]

[0049] $$\tilde{I}(x;\tilde{w})_{a_1} \geq \tilde{I}(x;\tilde{w})_{a_2} \geq \cdots \geq \tilde{I}(x;\tilde{w})_{a_n}$$

[0050] **(최적화)**

[0051] 그래디언트 왜곡을 위한 손실 함수(수학식 1)와 쿼리 출력 품질 보존을 위한 제약 조건(수학식 2, 수학식 3)을

결합하여 왜곡 모델을 생성하기 위한 COP(constrained optimization problem, 제약 조건이 있는 최적화 문제)를 다음과 같이 구성할 수 있다. 아래 수학식은 최적화 문제의 목적 함수(objective function)와 제약 조건을 나타낸다.

[0052]
$$\min_{\tilde{w}} L_{\text{orth}}(\tilde{w})$$
$$s.t.\ \tilde{f}(x;\tilde{w})_{s_1} \geq \cdots \geq \tilde{f}(x;\tilde{w})_{s_k} \geq \max_{j \in S'} \tilde{f}(x;\tilde{w})_j$$
$$\tilde{I}(x;\tilde{w})_{a_1} \geq \tilde{I}(x;\tilde{w})_{a_2} \geq \cdots \geq \tilde{I}(x;\tilde{w})_{a_n}$$

[0053] 하지만, 쿼리 출력 품질 보존을 위한 제약 조건을 나타내는 함수의 비선형성으로 인해 해당 COP(constrained optimization problem)를 직접 푸는 것은 매우 어렵다. 따라서, 본 발명에서는 COP 문제를 직접 풀지 않고 벌칙 기법(penalty method)을 적용하여 제약 조건을 갖지 않는 최적화 문제(unconstrained optimization problem)로 변환한 후 왜곡 모델을 생성한다.

[0054] 본 발명에서는, 도 3 및 도 4에 도시된 바와 같이, 주어진 쿼리 입력 x에 대해 매번 제약 조건을 갖지 않는 최적화 문제를 풀이하여 왜곡 모델을 새로 생성하여 왜곡 모델의 쿼리 출력을 노출 시킴으로써 모델 탈취 공격을 저지할 수 있다.

[0055] **(계산 효율성 증대를 위한 레이어 선택 전략)**

[0056] 인공지능(Artificial Intelligence, AI) 모델을 구성하는 모든 레이어(layer)를 그래디언트 왜곡에 참여 시킬 경우 연산 시간이 증가한다는 문제점 존재한다. 따라서, 본 발명에서는 계산 효율성을 높이기 위해서 모델의 출력에 대한 민감도가 높은 레이어를 선택적으로 그래디언트 왜곡에 참여시킨다. 레이어별 민감도는 다음과 같이 정의된다. 아래 식에서 ℓ은 레이어의 인덱스, w_ℓ은 ℓ 번째 레이어의 학습 파라미터를 나타낸다.

[0057]
$$S_\ell := \frac{1}{N}\sum_{i=1}^{N} \|\nabla_{w_\ell} f(x_i;w)_{y_i}\|_1$$

[0058] 레이어별 민감도를 내림차순으로 정렬 했을 때, (i)를 i 번째로 민감도가 큰 레이어의 인덱스로 정의한다($S_{(1)} \geq S_{(2)} \geq \cdots \geq S_{(L)}$). 여기서, L은 총 레이어의 개수로써, 1 보다 크거나 같은 자연수이다. 이때, 누적 민감도를 다음과 같이 정의한다.

[0059]
$$CS(\ell) := \frac{\sum_{i=1}^{\ell} S_{(i)}}{\sum_{i=1}^{L} S_{(i)}} \times 100(\%)$$

[0060] 본 발명에서는 누적 민감도를 그래디언트 왜곡에 참여시킬 레이어의 수를 판단하는 지표로 활용한다. 실시예에 따르면, 누적 민감도가 m(m은 0 보다 크거나 같고 100 보다 작거나 같은 임의의 실수로써, 예시적인 값은 90 임)%가 되는 최소 개수의 레이어를 그래디언트 왜곡에 참여시킬 수 있다. 그러나, 본 발명이 이에 제한되는 것은 아니며, 더 높은 계산 효율성을 위해서 더 낮은 누적 민감도가 달성되도록 조정이 가능하다. 즉, 그래디언트 왜곡에 참여하는 레이어는 민감도가 높은 적어도 하나의 레이어로 이해될 수 있다.

[0061] **(해석을 동반하지 않는 경우의 적용 방안)**

[0062] 상술한 제안 기법은 해석을 포함하지 않는 분류 시스템에도 적용이 가능하다. 해석을 포함하지 않는 경우, 왜곡 모델을 생성하는 최적화 식이 다음과 같이 변형된다. 즉, 해석 품질 보존을 위한 제약 조건이 사라진 형태를 갖는다.

[0063]
$$\min_{\tilde{w}} L_{\text{orth}}(\tilde{w})$$
$$s.t.\ \tilde{f}(x;\tilde{w})_{s_1} \geq \cdots \geq \tilde{f}(x;\tilde{w})_{s_k} \geq \max_{j \in S'} \tilde{f}(x;\tilde{w})_j$$

[0064] 해석을 동반하는 경우와 마찬가지로, 해당 문제를 직접 풀지 않고 벌칙 기법(penalty method)을 적용하여 제약 조건을 갖지 않는 최적화 문제로 변환한 후 왜곡 모델을 생성한다.

[0065] 주어진 쿼리 입력 x에 대해 매번 제약 조건을 갖지 않는 최적화 문제를 풀이하여 왜곡 모델을 새로 생성하고,

왜곡 모델의 쿼리 출력을 노출 시킴으로써 모델 탈취 공격을 저지할 수 있다.

[0066] 도 5는 본 발명의 일 실시예에 따른 모델 탈취 방어 방법의 개요도이고, 도 6은 본 발명의 일 실시예에 따른 모델 탈취 방어 방법의 흐름도이다.

[0067] 모델 탈취 방어 방법은 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치로 구현되는 모델 탈취 방어 장치에 의해 수행된다. 따라서, 모델 탈취 방어 방법에 포함되는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 랩탑 컴퓨터, 태블릿 PC 등을 포함한다. 또한, 모델 탈취 방어 방법을 설명함에 있어 앞선 기재와 중복되는 내용에 관하여는 그 구체적인 기재를 생략하기로 한다.

[0068] 우선, 쿼리 입력이 수신된다(S110). 쿼리 입력은 소정의 유무선 통신망을 통하여 사용자 단말로부터 수신될 수 있다. 즉, 사용자는 쿼리 입력에 대한 분류 결과를 얻기 위하여 쿼리 입력을 송신한다. 실시예에 의하면, 쿼리 입력은 분류 대상이 되는 이미지를 의미할 수 있다.

[0069] 다음으로, 분류 모델에 대한 왜곡 모델이 생성된다(S120). 분류 모델은 쿼리 입력에 대한 분류 결과를 도출하기 위해 학습된 인공 신경망 모델을 의미할 수 있다. 이때, 분류 모델은 분류 결과에 대한 해석을 동반하는 분류 모델이거나 해석을 동반하지 않는 분류 모델일 수 있다. 각 분류 모델에 대한 왜곡 모델의 생성 방법에 대한 구체적인 설명은 생략하기로 한다.

[0070] 왜곡 모델의 쿼리 출력이 생성된다(S130). 쿼리 출력은 분류 결과 및/또는 분류 결과에 대한 해석을 포함할 수 있다.

[0071] 마지막으로, 쿼리 출력은 사용자 단말로 송신된다(S140). 따라서, 사용자는 송신한 쿼리 입력에 대한 분류 결과 및/또는 설명을 획득할 수 있다.

[0072] 상술한 본 발명의 실시예들에 의할 경우, 다음과 같은 효과를 얻을 수 있다.

[0073] 1) 딥러닝 기반 분류 모델의 분류 결과에 대한 신뢰성과 해석가능성을 확보하기 위해서는 XAI 기법을 통해 생성한 분류 결과 설명을 쿼리 출력에 포함시켜야 한다. 본 발명에서 제공하는 그래디언트 왜곡(gradient misdirection) 기반 모델 탈취 방어 기술을 통해 쿼리 출력을 구성하는 분류 결과와 분류 결과에 대한 해석(interpretation)을 모두 보호하여 모델 탈취 공격에 대한 취약성을 감소시킬 수 있다.

[0074] 2) 본 발명에서 제안하는 모델 탈취 방어 기술은 모델 탈취 공격자가 사용하는 모델이나 데이터셋 유형에 대한 어떠한 가정도 하고 있지 않다. 따라서 다양한 유형의 모델 탈취 공격자에 대해서 균일한 방어 성능을 보일 수 있으며 공격자가 방어 기술의 전제조건을 악용하여 적응형 공격(adaptive attack)에도 높은 안정성을 보일 수 있다.

[0075] 3) 본 발명에서 제안하는 모델 탈취 방어 기술은 높은 모델 탈취 방어 성능을 보이는 동시에 타겟 모델의 쿼리 출력을 구성하는 분류 결과의 정확도와 해석의 품질을 보존하기 위한 메커니즘을 포함하고 있다. 본 발명을 통해 분류 모델의 분류 성능과 해석가능성에 대한 희생 없이 높은 모델 탈취 방어 성능을 달성하는 것이 가능하다.

[0076] 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 사용되는 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.

[0077] 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

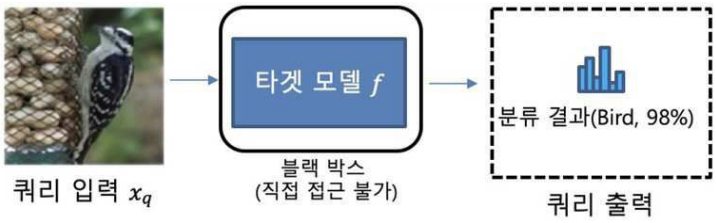
(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0078] 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.

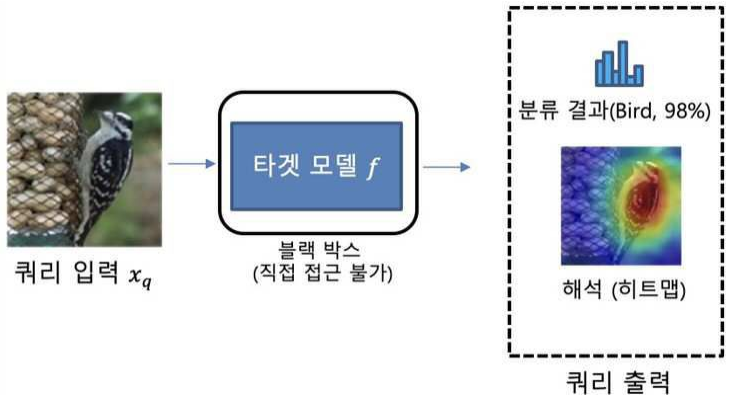
[0079] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

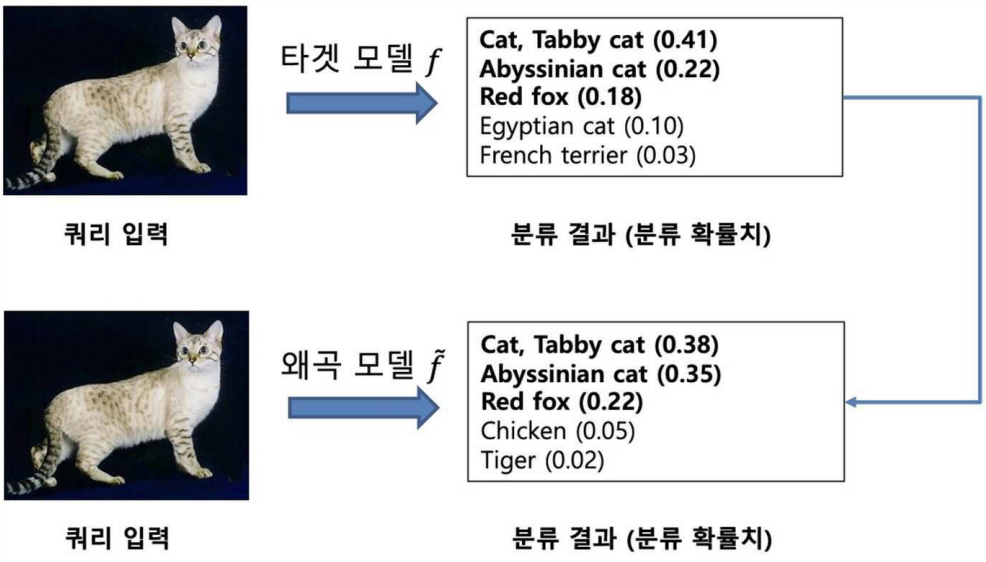
도면1



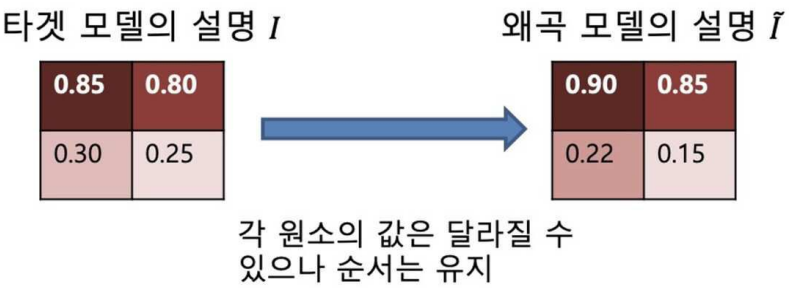
도면2



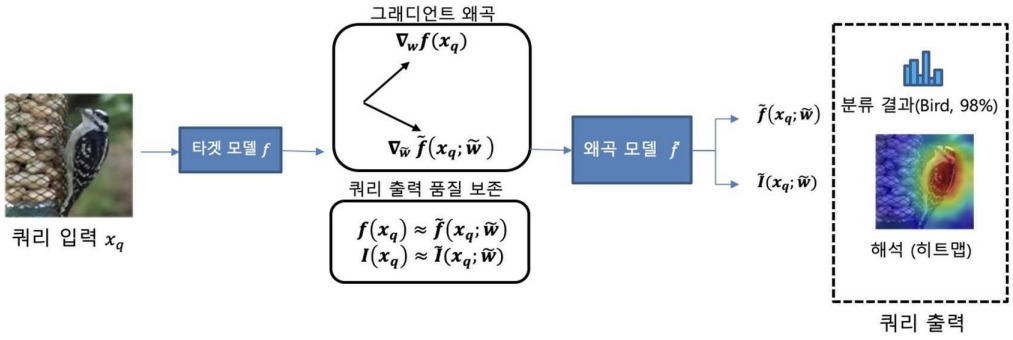
도면3


타겟 모델 f
Cat, Tabby cat (0.41)
Abyssinian cat (0.22)
Red fox (0.18)
Egyptian cat (0.10)
French terrier (0.03)
쿼리 입력
분류 결과 (분류 확률치)
왜곡 모델 $\tilde{f}$
Cat, Tabby cat (0.38)
Abyssinian cat (0.35)
Red fox (0.22)
Chicken (0.05)
Tiger (0.02)
쿼리 입력
분류 결과 (분류 확률치)

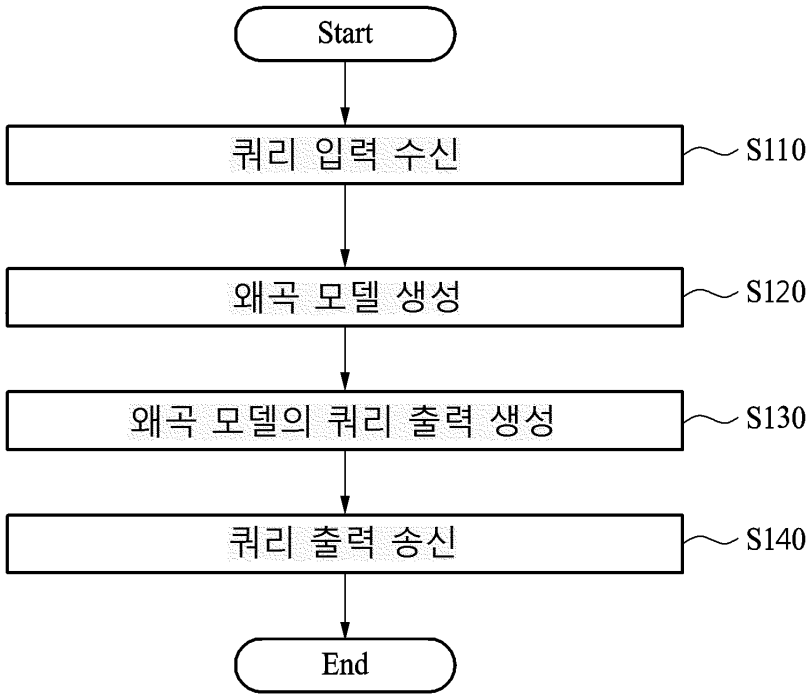
도면4


타겟 모델의 설명 I
왜곡 모델의 설명 $\tilde{I}$
0.85
0.80
0.30
0.25
0.90
0.85
0.22
0.15
각 원소의 값은 달라질 수
있으나 순서는 유지

도면5


그래디언트 왜곡
$\nabla_w f(x_q)$
$\nabla_{\tilde{w}} \tilde{f}(x_q; \tilde{w})$
쿼리 출력 품질 보존
$f(x_q) \approx \tilde{f}(x_q; \tilde{w})$
$I(x_q) \approx \tilde{I}(x_q; \tilde{w})$
타겟 모델 f
왜곡 모델 $\tilde{f}$
$\tilde{f}(x_q; \tilde{w})$
$\tilde{I}(x_q; \tilde{w})$
분류 결과(Bird, 98%)
해석 (히트맵)
쿼리 입력 x_q
쿼리 출력

도면6


Start
쿼리 입력 수신
S110
왜곡 모델 생성
S120
왜곡 모델의 쿼리 출력 생성
S130
쿼리 출력 송신
S140
End