



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0033382
(43) 공개일자 2025년03월10일

(51) 국제특허분류(Int. Cl.)

G06V 10/82 (2022.01) G06V 10/26 (2022.01)
G06V 10/764 (2022.01) G06V 10/776 (2022.01)

(52) CPC특허분류

G06V 10/82 (2022.01)
G06N 3/08 (2023.01)

(21) 출원번호 10-2023-0113066

(22) 출원일자 2023년08월28일

심사청구일자 2023년08월28일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

이성환

서울특별시 강남구 언주로30길 13 (대림아크로빌, 대림아크로텔) B-2902

전규성

서울특별시 마포구 만리재로14길 53 (공덕동) 옥탑방

(뒷면에 계속)

(74) 대리인

특허법인위솔

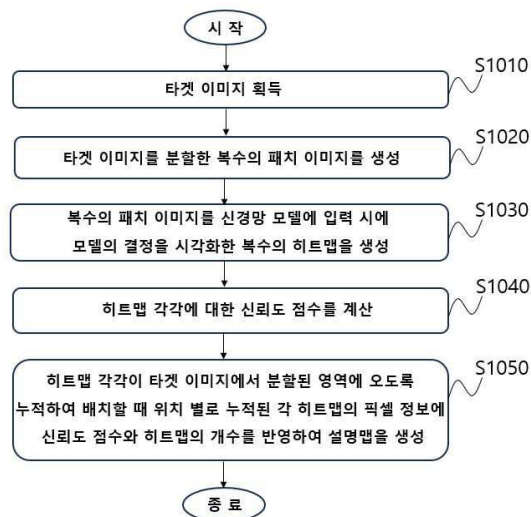
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 분할 및 취합 귀속 지침을 이용한 신경망 결정의 시각화 방법 및 이를 위한 장치

(57) 요약

일 실시 예에 따른 시각화 장치는 타겟 이미지를 획득하는 동작; 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작; 상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작; 상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및 상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작을 수행할 수 있다.

대표도 - 도2



(52) CPC특허분류

G06N 5/045 (2023.01)

G06V 10/26 (2023.08)

G06V 10/764 (2023.08)

G06V 10/776 (2023.08)

G06T 2219/2016 (2013.01)

(72) 발명자

남우정

서울특별시 동대문구 무학로 193, 404호 (용두동, 삼영 오피스텔)

홍중호

경기도 수원시 장안구 정자로42번길 52 (천천동 비단마을 베스트타운) 731동 1203호

이 발명을 지원한 국가연구개발사업

과제고유번호 1711193255

과제번호 2022-0-00984-002

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 사람중심인공지능핵심원천기술개발

연구과제명 플러그앤플레이 방식으로 설명가능성을 제공하는 인공지능 기술 개발 및 인공지능 시스템에 대한 설명 제공 검증

기 여 율 1/2

과제수행기관명 한국과학기술원

연구기간 2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711193973

과제번호 2019-0-00079-005

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 정보통신방송혁신인재양성

연구과제명 인공지능대학원지원(고려대학교)

기 여 율 1/2

과제수행기관명 고려대학교산학협력단

연구기간 2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

프로세서 및 메모리를 포함하는 시각화 장치가 수행하는 시각화 방법에 있어서,

타겟 이미지를 획득하는 동작;

타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작;

상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작;

상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및

상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작;

을 포함하는 방법.

청구항 2

제1항에 있어서,

상기 복수의 패치 이미지를 생성하는 동작은,

상기 타겟 이미지 보다 작은 기 설정된 크기의 윈도우를 생성하는 동작;

상기 윈도우가 상기 타겟 이미지의 모든 영역을 지나도록 기 설정된 간격으로 상기 윈도우를 이동시키는 동작; 및

상기 간격 마다 상기 윈도우에 포함되는 타겟 이미지의 영역을 캡처하여 상기 복수의 패치 이미지를 생성하는 동작을 포함하는,

방법.

청구항 3

제1항에 있어서,

상기 복수의 패치 이미지를 생성하는 동작은,

상기 타겟 이미지로부터 각 패치 이미지가 분할된 위치 정보를 저장하는 동작을 포함하는,

방법.

청구항 4

제1항에 있어서,

상기 복수의 패치 이미지를 생성하는 동작은,

상기 복수의 패치 이미지를 상기 타겟 이미지의 크기로 리사이징 하는 동작을 포함하고,

상기 설명맵을 생성하는 동작은,

상기 복수의 히트맵 각각을 상기 일부 영역의 크기로 리사이징 하여, 상기 일부 영역의 크기로 리사이징된 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 위치하도록 배치하는 동작을 포함하는,

방법.

청구항 5

제1항에 있어서,

상기 복수의 히트맵을 생성하는 동작은,

상기 신경망 모델에 패치 이미지가 입력되는 경우 상기 신경망 모델에 클래스 활성화 매핑(Class Activation Mapping) 기법 또는 계층별 관련성 전파(Layer-wise Relevance Propagation) 기법을 적용하여, 상기 신경망 모델의 결정에 대한 정보를 시각화한 히트맵을 생성하는 동작을 포함하는,

방법.

청구항 6

제1항에 있어서,

상기 신뢰도 점수를 계산하는 동작은,

상기 신경망 모델의 모든 클래스의 로짓값에 대한 패치 이미지에서 관찰하려는 클래스를 기준으로 상기 신경망 모델이 출력한 로짓값의 비율을 기초로 패치 이미지에 대한 히트맵의 신뢰도 점수를 계산하는 동작을 포함하는,

방법.

청구항 7

제1항에 있어서,

상기 신뢰도 점수를 판별하는 동작은,

상기 신경망 모델에 패치 이미지를 입력하는 경우 출력되는 로짓값에 하기 수학적 식 1을 적용하여 패치 이미지 별 신뢰도 점수를 계산하는,

[수학적 식 1]

$$Y_s^k = \xi(\mathcal{L}_s, k) \\ = \exp \left(\frac{\exp(\mathcal{L}_s^{k,j})}{\sum_{i=1}^N \exp(\mathcal{L}_s^{i,j})} \right), \forall j$$

(Y_s^k 는 신뢰도 점수, $\mathcal{L}_s$ 는 패치 이미지에 대한 로짓값, k 는 패치 이미지에서 관찰하려는 클래스, $\xi(\cdot)$ 는 $\mathcal{L}_s$ 와 k 를 입력으로 하여 신뢰도 점수를 계산하는 함수, N 은 클래스의 총 개수, i 는 클래스를 특정하는 인덱스, j 는 로짓을 출력하는 최종 노드를 특정하는 인덱스)

방법.

청구항 8

제1항에 있어서,

상기 설명맵을 생성하는 동작은,

제1 위치에 누적된 각 히트맵의 픽셀 정보에 각 히트맵의 신뢰도 점수를 가중치로 가하여 합산 정보를 생성하고, 상기 제1 위치에 누적된 히트맵의 개수로 상기 합산 정보를 정규화하여 설명맵을 생성하는 동작을 포함하는,

방법.

청구항 9

명령어를 포함하는 메모리; 및

상기 명령어를 실행하는 프로세서를 포함하고,

상기 프로세서의 동작은,

타겟 이미지를 획득하는 동작;

타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작;

상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작;

상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및

상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작;

을 포함하는 장치.

청구항 10

컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은, 프로세서에 의해 실행되면,

타겟 이미지를 획득하는 동작;

타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작;

상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작;

상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및

상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작;

을 포함하는 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 분할 및 취합 귀속 지침을 이용한 신경망 결정의 시각화 방법 및 이를 위한 장치에 관한 것이다. 보다 자세하게는 타겟 이미지의 분할과 시각화 정보의 취합 방식을 통해 신경망의 결정을 시각화하는 방법 및 이를 위한 장치에 관한 것이다.

배경 기술

[0002] 인공지능 신경망을 이용한 이미지 분류 신경망 모델은 특정 이미지와 클래스 레이블이 매핑된 데이터의 집합을 이용하여, 이미지로부터 정답 클래스를 분류하도록 학습된다.

[0003] 이때, 신경망 모델의 입력과 출력 사이는 블랙 박스 형태로, 신경망 모델이 입력으로부터 출력까지의 결정을 어떠한 과정 또는 어떠한 방법으로 처리하는 것인지 직접적으로는 알기가 어렵다. 그렇기 때문에, 신경망 모델이 특정한 결정을 내리는 이유를 파악할 수 있다면, 인공지능을 보다 이해하여 개선할 수 있는 가능성이 열리므로, 신경망 모델의 처리 과정을 시각화하여 인간이 이해할 수 있도록 하는 ‘설명 가능한 인공지능’의 연구가 활발하다.

[0004] 한편, 이미지 분류 신경망 모델의 경우, 신경망 모델이 입력된 이미지에 대하여 특정한 클래스를 결정하게 되는 이유를 파악하기 위해, 신경망 모델이 정답을 고를 때 가장 많이 참고한 부분을 시각화하려는 시도가 존재하였

다.

- [0005] 예를 들어, 클래스 활성화 매핑(Class Activation Mapping)과 계층별 관련성 전파(Layer-wise Relevance Propagation) 기술은 신경망 모델이 입력된 이미지의 어떤 영역을 주로 참조하여 결과를 도출하였는지 시각화하기 위한 방법으로 사용되고 있다.
- [0006] 한편, 클래스 활성화 매핑 기술의 경우, 입력 이미지와 시각화 설명맵의 해상도 차이로 인하여 설명맵이 표현할 수 있는 범위가 제한된다. 예를 들어, CNN의 ResNet50의 경우에는 224 픽셀 단위의 정사각형 이미지가 입력되는데, 설명맵은 7 픽셀 단위의 정사각형 이미지로 표현되어, 32배의 해상도 차이가 발생하여 표현력이 제한된다는 한계가 있다.
- [0007] 또한, 계층별 관련성 전파 기술은 픽셀 단위로 표현 시 이미지에 노이즈가 발생하여 표현이 불명확해지는 현상이 존재하고, 특정 클래스에 특화된 설명을 제공하는 능력이 부족하여 설명맵의 신뢰도가 낮다는 한계가 있다.
- [0008] 따라서, 상술한 한계점을 극복할 수 있는 신경망 모델의 결정 시각화 방법론이 개발된다면, 인공지능 신경망의 이해와 발전에 큰 도움이 될 것으로 예상된다.

선행기술문헌

특허문헌

- [0009] (특허문헌 0001) 대한민국 등록특허공보 제10-2446792호(2022.09.20)

발명의 내용

해결하려는 과제

- [0010] 본 발명이 해결하고자 하는 기술적 과제는 신경망 모델의 구조와 무관하게 설명맵의 세분성과 클래스별 명시적 정확성을 높이는 기술을 제공하는 것이다. 이를 위해, 타겟 이미지를 분할하여 타겟 이미지의 부분적 정보를 포함하는 패치 이미지를 생성하고, 패치 이미지에 대한 히트맵의 시각화 정보를 취합하는 방식의 접근을 통해, 설명맵의 수준을 높여 보다 향상된 '설명 가능한 인공지능(Explainable Artificial Intelligence)' 기술을 제공하고자 한다.
- [0011] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

- [0012] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 프로세서 및 메모리를 포함하는 시각화 장치가 수행하는 시각화 방법은, 타겟 이미지를 획득하는 동작; 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작; 상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작; 상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및 상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작을 포함할 수 있다.
- [0013] 일 실시 예에 따르면, 상기 복수의 패치 이미지를 생성하는 동작은 상기 타겟 이미지 보다 작은 기 설정된 크기의 윈도우를 생성하는 동작; 상기 윈도우가 상기 타겟 이미지의 모든 영역을 지나도록 기 설정된 간격으로 상기 윈도우를 이동시키는 동작; 및 상기 간격 마다 상기 윈도우에 포함되는 타겟 이미지의 영역을 캡처하여 상기 복수의 패치 이미지를 생성하는 동작을 포함할 수 있다.
- [0014] 일 실시 예에 따르면, 상기 복수의 패치 이미지를 생성하는 동작은 상기 타겟 이미지로부터 각 패치 이미지가 분할된 위치 정보를 저장하는 동작을 포함할 수 있다.
- [0015] 일 실시 예에 따르면, 상기 복수의 패치 이미지를 생성하는 동작은 상기 복수의 패치 이미지를 상기 타겟 이미지의 크기로 리사이징 하는 동작을 포함하고,

[0016] 상기 설명맵을 생성하는 동작은 상기 복수의 히트맵 각각을 상기 일부 영역의 크기로 리사이징 하여, 상기 일부 영역의 크기로 리사이징된 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 위치하도록 배치하는 동작을 포함할 수 있다.

[0017] 일 실시 예에 따르면, 상기 복수의 히트맵을 생성하는 동작은 상기 신경망 모델에 패치 이미지가 입력되는 경우 상기 신경망 모델에 클래스 활성화 매핑(Class Activation Mapping) 기법 또는 계층별 관련성 전파(Layer-wise Relevance Propagation) 기법을 적용하여, 상기 신경망 모델의 결정에 대한 정보를 시각화한 히트맵을 생성하는 동작을 포함할 수 있다.

[0018] 일 실시 예에 따르면, 상기 신뢰도 점수를 계산하는 동작은 상기 신경망 모델의 모든 클래스의 로짓값에 대한 패치 이미지에서 관찰하려는 클래스를 기준으로 상기 신경망 모델이 출력한 로짓값의 비율을 기초로 패치 이미지에 대한 히트맵의 신뢰도 점수를 계산하는 동작을 포함할 수 있다.

[0019] 일 실시 예에 따르면, 상기 신뢰도 점수를 판별하는 동작은 상기 신경망 모델에 패치 이미지를 입력하는 경우 출력되는 로짓값에 하기 수학적 1을 적용하여 패치 이미지 별 신뢰도 점수를 계산할 수 있다.

[0020] [수학적 1]

$$Y_s^k = \xi(\mathcal{L}_s, k)$$

$$= \exp \left(\frac{\exp(\mathcal{L}_s^{k,j})}{\sum_{i=1}^N \exp(\mathcal{L}_s^{i,j})} \right), \forall j$$

[0021]

[0022] (Y_s^k 는 신뢰도 점수, $\mathcal{L}_s$ 는 패치 이미지에 대한 로짓값, k 는 패치 이미지에서 관찰하려는 클래스, $\xi(\cdot)$ 는 $\mathcal{L}_s$ 와 k 를 입력으로 하여 신뢰도 점수를 계산하는 함수, N 은 클래스의 총 개수, i 는 클래스를 특정하는 인덱스, j 는 로짓을 출력하는 최종 노드를 특정하는 인덱스)

[0023] 일 실시 예에 따르면, 상기 설명맵을 생성하는 동작은 제1 위치에 누적된 각 히트맵의 픽셀 정보에 각 히트맵의 신뢰도 점수를 가중치로 가하여 합산 정보를 생성하고, 상기 제1 위치에 누적된 히트맵의 개수로 상기 합산 정보를 정규화하여 설명맵을 생성하는 동작을 포함할 수 있다.

[0024] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 시각화 장치는 명령어를 포함하는 메모리; 및 상기 명령어를 실행하는 프로세서를 포함하고, 상기 프로세서의 동작은 타겟 이미지를 획득하는 동작; 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작; 상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작; 상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및 상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작을 포함할 수 있다.

[0025] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램으로서, 상기 컴퓨터 프로그램은, 프로세서에 의해 실행되면, 타겟 이미지를 획득하는 동작; 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작; 상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작; 상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및 상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작을 포함하는 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함할 수 있다.

발명의 효과

[0026] 상기와 같은 본 발명에 따르면, 타겟 이미지를 분할하여 타겟 이미지의 부분적 정보를 포함하는 패치 이미지를 생성하고, 패치 이미지에 대한 히트맵의 시각화 정보를 취합하여, 세분성과 정확성이 강화된 설명법을 도출함으로써, 신경망 모델이 타겟 이미지의 어떤 영역을 주로 참고하여 특정한 결정을 내리는지 시각화 할 수 있다.

[0027] 이에 따라, 본 발명은 타겟 이미지에서 어떠한 특징이 신경망 모델의 의사 결정에 주요한 영향을 미치는지 직관적으로 확인할 수 있게 한다. 특히, 의학 분야와 같이 의사 결정의 부담이 큰 분야에서는, 진단의 근거를 제시하여 진단 과정의 정확성과 효율을 향상시킬 수 있으며, 자율 주행과 같이 인명 피해와 직접적으로 연관된 분야에서는 신경망 모델의 의사 결정이 잘못된 특징으로부터 유도되는지 미리 확인하여 수정할 수 있게 한다.

[0028] 이처럼, 본 발명을 다양한 분야에 활용하는 경우, 사용자는 신경망 모델 및 데이터셋 개선을 위한 방향성을 제공받아 신경망 모델을 보다 효과적으로 활용할 수 있으며, 나아가 인공지능 신경망 분야의 이해와 발전에 기여할 수 있다.

[0029] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0030] 도 1은 일 실시 예에 따른 시각화 장치가 포함하는 전체 구성을 예시적으로 도시한 도면이다.

도 2는 일 실시 예에 따른 시각화 장치가 수행하는 시각화 방법의 대표적인 단계를 나타낸 순서도이다.

도 3은 도 2의 실시 예에 따른 시각화 방법의 각 동작에 대한 이해를 돕기 위한 예시도이다.

도 4는 일 실시 예에 따라 S1020 단계의 패치 이미지를 생성하는 동작을 예시적으로 구체화한 순서도이다.

도 5는 도 4의 실시 예에 따른 패치 이미지를 생성 단계의 각 동작에 대한 이해를 돕기 위한 예시도이다.

도 6은 일 실시 예에 따라 설명맵을 생성하는 동작에 대한 이해를 돕기 위한 예시도이다.

도 7은 기존의 시각화 방법을 적용하여 설명맵을 생성한 결과와, 기존의 시각화 방법에 본 발명의 실시 예를 적용하여 설명맵을 생성한 결과를 비교한 도면이다.

도 8은 ResNet50과 VGG19의 신경망 모델에 대해, 기존의 시각화 방법을 적용하여 설명맵을 생성한 결과와, 기존의 시각화 방법에 본 발명의 실시 예를 적용하여 설명맵을 생성한 결과를 비교한 도면이다.

도 9는 ResNet50, DenseNet121, InceptionV3의 신경망 모델에 기존의 시각화 방법(GCAM, GCAM++, WGCAM)을 적용한 경우 deletion/insertion game test의 성능과, 기존의 시각화 방법에 본 발명의 실시 예를 적용한 경우((GCAM*, GCAM++*, WGCAM*)) deletion/insertion game test 성능을 비교한 성능 평가 테이블이다.

도 10은 다양한 시각화 설명 기법과 본 발명의 실시 예를 적용한 경우, 각 설명맵이 객체를 특정하는 정도에 대한 점수를 비교한 성능 평가 테이블이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명의 목적과 기술적 구성 및 그에 따른 작용 효과에 관한 자세한 사항은 본 발명의 명세서에 첨부된 도면에 의거한 이하의 상세한 설명에 의해 보다 명확하게 이해될 것이다. 첨부된 도면을 참조하여 본 발명에 따른 실시 예를 상세하게 설명한다.

[0032] 본 명세서에서 개시되는 실시 예들은 본 발명의 범위를 한정하는 것으로 해석되거나 이용되지 않아야 할 것이다. 이 분야의 통상의 기술자에게 본 명세서의 실시 예를 포함한 설명은 다양한 응용을 갖는다는 것이 당연하다. 따라서, 본 발명의 상세한 설명에 기재된 임의의 실시 예들은 본 발명을 보다 잘 설명하기 위한 예시적인 것이며 본 발명의 범위가 실시 예들로 한정되는 것을 의도하지 않는다.

[0033] 도면에 표시되고 아래에 설명되는 기능 블록들은 가능한 구현의 예들일 뿐이다. 다른 구현들에서는 상세한 설명의 사상 및 범위를 벗어나지 않는 범위에서 다른 기능 블록들이 사용될 수 있다. 또한, 본 발명의 하나 이상의 기능 블록이 개별 블록들로 표시되지만, 본 발명의 기능 블록들 중 하나 이상은 동일 기능을 실행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합일 수 있다.

[0034] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.

[0035] 나아가 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급될 때에는, 그 다른 구성요소에 직접적으로 연결 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 한다.

- [0036] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.
- [0037] 도 1은 일 실시 예에 따른 시각화 장치(100)(이하, '장치(100)'로 지칭)가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- [0038] 그러나 이는 본 발명의 목적을 달성하기 위한 바람직한 실시 예일 뿐이며, 필요에 따라 일부 구성이 추가되거나 삭제될 수 있고, 어느 한 구성이 수행하는 역할을 다른 구성이 함께 수행할 수도 있음은 물론이다.
- [0039] 일 실시 예에 따른 장치(100)는 프로세서(10), 네트워크 인터페이스(20), 메모리(30), 스토리지(40) 및 이들을 연결하는 데이터 버스(50)를 포함할 수 있으며, 기타 본 발명의 목적을 달성함에 있어 요구되는 부가적인 구성들을 더 포함할 수 있음은 물론이라 할 것이다.
- [0040] 프로세서(10)는 각 구성의 전반적인 동작을 제어한다. 프로세서(10)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit) 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 형태의 인공지능 프로세서 중 어느 하나일 수 있다. 아울러, 프로세서(10)는 일 실시 예에 따른 장치(100)가 수행하는 동작 또는 오퍼레이션을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.
- [0041] 네트워크 인터페이스(20)는 일 실시 예에 따른 장치(100)의 유무선 인터넷 통신을 지원하며, 그 밖의 공지의 통신 방식을 지원할 수도 있다. 따라서 네트워크 인터페이스(20)는 그에 따른 통신 모듈을 포함하여 구성될 수 있다.
- [0042] 메모리(30)는 각종 정보, 명령 및/또는 정보를 저장하며, 일 실시 예에 따른 장치(100)가 수행하는 동작 또는 오퍼레이션을 수행하기 위해 스토리지(40)로부터 하나 이상의 컴퓨터 프로그램(41)을 로드할 수 있다. 도 1에서는 메모리(30)의 하나로 RAM을 도시하였으나 이와 더불어 다양한 저장 매체를 메모리(30)로 이용할 수 있음은 물론이다.
- [0043] 스토리지(40)는 하나 이상의 컴퓨터 프로그램(41)을 비임시적으로 저장할 수 있다. 이러한 스토리지(40)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크(HDD), 보조 저장 매체(SSD), 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체 중 어느 하나일 수 있다.
- [0044] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해, 타겟 이미지를 획득하는 동작; 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작; 상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작; 상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및 상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작을 실행할 수 있다.
- [0045] 이상 간단하게 언급한 컴퓨터 프로그램(41)이 수행하는 동작 또는 오퍼레이션은 컴퓨터 프로그램(41)의 일 기능으로 볼 수 있으며, 보다 자세한 설명은 일 실시 예에 따른 장치(100)가 수행하는 방법에 대한 설명에서 후술하도록 한다.
- [0046] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의 명령 및/또는 정보의 이동 경로가 된다.
- [0047] 이상 간단하게 설명한 일 실시 예에 따른 장치(100)는 독립된 디바이스의 형태, 예를 들어 전자 기기나 서버(클라우드 포함)의 형태일 수 있으며, 여기서 전자 기기는 한 장소에 고정 설치되어 사용하는 데스크톱 PC, 서버 디바이스 등과 같은 기기 뿐만 아니라, 스마트폰, 태블릿 PC, 노트북 PC, PDA, PMP 등과 같이 휴대가 용이한 포터블 기기 등이라도 무방한바, 프로세서(10)에 해당하는 CPU 등이 설치되고 네트워크 기능만 보유하고 있는 전자 기기라면 어떠한 것이라도 무방하다 할 것이다. 예를 들어, 일 실시 예에 따른 장치(100)는 독립된 디바이스인 전자 기기 형태 중 "서버"의 형태로 구현될 수 있다. 또한, 일 실시 예에 따른 장치(100)는 독립된 디바이스인 전자 기기 형태 중 서버가 제공하는 애플리케이션 또는 서비스를 이용하기 위해 사용자가 소지하는 "단말"의 형태로 구현될 수 있다.
- [0048] 이하, 일 실시 예에 따른 장치(100)가 수행하는 방법에 대하여 도 2 내지 도 10을 참조하여 설명하도록 한다.

- [0049] 도 2는 일 실시 예에 따라 장치(100)가 수행하는 시각화 방법의 대표적인 단계를 나타낸 순서도이고, 도 3은 도 2의 실시 예에 따른 시각화 방법의 각 단계에 대한 이해를 돕기 위해 예시한 개념도이다.
- [0050] 도 2 및 도 3에 개시된 각 단계는 일 실시 예에 따른 장치(100)를 통해 수행될 수 있고, 구체적으로 각 단계의 동작은 프로세서(10)가 수행하는 동작으로 이해될 수 있다.
- [0051] 도 2 및 도 3에 개시된 각 단계는 본 발명의 목적을 달성함에 있어서 바람직한 실시 예일 뿐이며, 필요에 따라 일부 단계가 추가 또는 삭제될 수 있음은 물론이고, 어느 한 단계가 다른 단계에 포함되어 수행될 수도 있다.
- [0052] 도 2 및 도 3에 개시된 각 동작의 순서는 이해의 편의를 위해 배치된 순서일 뿐, 이러한 순서가 시계열적인 순서로 한정되는 것이 아니며, 설계자의 선택에 따라 순서가 다르게 변경되어 동작될 수 있다.
- [0053] 도 2 및 도 3을 함께 참조하면, S1010 단계에서, 장치(100)는 타겟 이미지를 획득할 수 있다. 여기서, 타겟 이미지는 사용자가 관찰하고자 하는 신경망 모델에 입력되는 이미지를 의미한다.
- [0054] 또한, S1010 단계에서, 장치(100)는 '설명 가능한 인공지능'의 분석 대상인 신경망 모델을 획득할 수 있다. 신경망 모델은 타겟 이미지를 입력 받아 타겟 이미지의 클래스를 결정하는 신경망 모델을 포함할 수 있다. 예를 들어, 신경망 모델은 다양한 학습 알고리즘을 기초로 이미지의 클래스를 분류하도록 학습된 신경망 모델을 포함할 수 있다.
- [0055] 본 문서의 실시 예에 따른 신경망 모델은 머신러닝에서 사용되는 모델로서, 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)들로 구성되는, 문제 해결 능력을 가지는 모델 전반을 의미한다.
- [0056] 신경망 모델은 입력 레이어(Input Layer), 출력 레이어(Output Layer), 그리고 선택적으로 하나 이상의 은닉 레이어(Hidden Layer)를 포함할 수 있다. 각 레이어는 하나 이상의 뉴런을 포함하고, 신경망 모델은 뉴런과 뉴런을 연결하는 시냅스를 포함할 수 있다. 신경망 모델에서 각 뉴런은 시냅스를 통해 입력되는 입력 신호들, 가중치, 편향에 대한 활성 함수의 함수값을 출력할 수 있다.
- [0057] 신경망 모델의 형태는 각 레이어의 뉴런들 사이의 연결 패턴, 모델 파라미터를 갱신하는 학습 과정, 출력값을 생성하는 활성화 함수(Activation Function)에 의해 정의될 수 있다. 신경망 모델의 각 레이어는 하나 이상의 뉴런을 포함할 수 있고, 뉴런과 뉴런을 연결하는 시냅스를 포함할 수 있다. 신경망 모델의 각 뉴런은 시냅스를 통해 입력되는 입력 신호들, 가중치, 편향에 대한 활성 함수의 활성화값을 출력할 수 있다.
- [0058] 신경망 모델의 모델 파라미터는 학습을 통해 결정되는 파라미터를 의미하며, 예를 들어 모델 파라미터는 시냅스 연결의 가중치와 뉴런의 편향 등을 포함할 수 있다. 또한, 하이퍼파라미터는 머신러닝 알고리즘에서 학습 전에 설정되어야 하는 파라미터를 의미하며, 하이퍼파라미터는 학습률(Learning Rate), 학습 횟수(epoch), 배치 크기(batch size), 초기화 함수 등을 포함할 수 있다.
- [0059] 이에, 본 발명의 실시 예는 소정의 신경망 모델에 대해 시각화 설명(Visualized explanation) 기법을 사용함에 있어서, 시각화 설명 기법의 전/후 단계에 분할과 취합 방식을 추가로 부가하여, 특정 시각화 설명 기법 자체에서 출력되는 설명맵의 수준을 높이는 방법을 제시한다. 즉, 본 발명은 어떠한 시각화 설명 기법을 사용하는 경우라도, 시각화 설명 기법의 전/후 단계에 앞으로 설명할 분할과 취합 방식을 접목하여 향상된 수준의 설명맵을 도출한다. 이후 단계들의 설명을 통해, 장치(100)가 시각화 설명 기법의 전/후 단계에서 타겟 이미지에 대한 분할과 시각화 정보를 취합하여 최종 설명맵을 생성하는 동작을 설명한다.
- [0060] 이에, S1020 단계에서, 장치(100)는 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성할 수 있다. 이때 장치(100)는 각 패치 이미지가 타겟 이미지에서 분할된 위치 정보를 저장할 수 있다.
- [0061] S1020 단계의 구체적인 예시를 다음 도 4 및 도 5와 함께 설명한다.
- [0062] 도 4는 일 실시 예에 따라 S1020 단계의 패치 이미지를 생성하는 동작을 예시적으로 구체화한 순서도이고, 도 5는 도 4의 실시 예에 따른 패치 이미지를 생성 단계의 각 동작에 대한 이해를 돕기 위한 예시도이다.
- [0063] 도 4 및 도 5를 참조하면, 장치(100)는 타겟 이미지 보다 작은 기 설정된 크기의 윈도우를 생성하고(S1021), 윈도우가 타겟 이미지의 모든 영역을 지나도록 기 설정된 간격으로 윈도우를 이동시켜(S1022), 윈도우의 이동 간격 마다 윈도우에 포함된 타겟 이미지의 영역을 캡처하여 복수의 패치 이미지를 생성할 수 있다(S1023). 여기서, 캡처된 패치 이미지는 윈도우 크기와 같은 크기로 타겟 이미지보다 작기 때문에, 장치(100)는 해상도의 통일을 위해 각 패치 이미지를 타겟 이미지의 크기와 동일한 크기가 되도록 리사이징하는 동작을 수행할 수 있다(S1024). 예를 들어, 장치(100)는 선형 보간법을 통해 패치 이미지를 타겟 이미지의 크기로 리사이징 할 수

있다.

[0064] 다음으로, S1030 단계에서, 장치(100)는 복수의 패치 이미지 각각을 관찰 대상인 신경망 모델에 입력 시에, 해당 신경망 모델의 결정을 시각화한 복수의 히트맵을 생성할 수 있다. 장치(100)는 다양한 시각화 설명 기법을 사용하여 신경망 모델의 결정을 시각화한 히트맵을 생성할 수 있다. 히트맵은 신경망 모델이 입력된 패치 이미지의 어느 영역을 주로 참조하여 클래스를 도출하였는지에 대한 정보가 시각화된 픽셀 정보(ex. GradCAM을 이용하여 생성한 도 3의 복수의 히트맵)를 포함할 수 있다.

[0065] 시각화 설명 기법은 신경망 모델이 입력된 이미지의 어느 영역을 주로 참조하여 클래스를 도출하였는지에 대한 정보를 시각화한 히트맵을 생성하는 기술이다. 예를 들어, 시각화 설명 기법은 클래스 활성화 매핑(Class Activation Mapping, 이하 'CAP' 로 지칭)과 계층별 관련성 전파(Layer-wise Relevance Propagation, 이하 'LRP' 로 지칭) 기술을 포함할 수 있다.

[0066] CAP는 기울기 기반 클래스 맵핑 기술(GradCAM)과 점수 기반 클래스 맵핑(ScoreCAM) 기술을 포함하며, GradCAM은 인공 신경망의 중간 계층의 특징과 특정 클래스에 대한 기울기를 선형 결합하여 시각화 설명을 제공하고, ScoreCAM은 중간 계층의 특징들을 직접적으로 입력 이미지를 교란, 모델의 예측율을 선형 결합하여 시각적 설명을 제공할 수 있다.

[0067] LRP는 신경망 모델의 역전파 방식을 재구성하여 픽셀 단위로 기여도를 계산하는 방식으로, 테일러 급수를 이용하여 신경망 모델의 결정을 근사할 수 있다.

[0068] 상술한 시각화 설명 기법의 종류들은 예시일 뿐, 실시 예의 적용 가능한 범위가 예시한 종류에 한정되는 것이 아니며, 본 발명의 실시 예에는 상술한 예시 이외에도 다양한 시각화 설명 기법을 접목할 수 있다.

[0069] 다음으로, S1040 단계에서, 장치(100)는 패치 이미지에 대한 신경망 모델의 로짓(logit)을 기초로 복수의 히트맵 각각에 대한 신뢰도 점수를 계산할 수 있다. 예를 들어, 장치(100) 신경망 모델의 모든 클래스의 로짓값에 대한 패치 이미지에서 관찰하려는 클래스를 기준으로 신경망 모델이 출력한 로짓값의 비율을 기초로 패치 이미지에 대한 히트맵의 신뢰도 점수를 계산할 수 있다. 여기서, 복수의 히트맵 중에는 타겟 이미지의 객체를 온전히 포함하지 않는 히트맵이 있을 가능성이 있기 때문에, 후술할 각 히트맵의 정보 취합 시 각각의 신뢰도 점수를 사용하게 된다.

[0070] 신뢰도 점수를 계산하는 실시 예는 보다 구체적으로 하기 수학적 식 1로 구현될 수 있다. 예를 들어, 장치(100)는 신경망 모델에 패치 이미지를 입력하는 경우 출력되는 로짓값에 하기 수학적 식 1을 적용하여 패치 이미지 별 신뢰도 점수를 계산할 수 있다.

[0071] [수학적 식 1]

$$Y_s^k = \xi(\mathcal{L}_s, k) \\ = \exp \left(\frac{\exp(\mathcal{L}_s^{k,j})}{\sum_{i=1}^N \exp(\mathcal{L}_s^{i,j})} \right), \forall j$$

[0072]

[0073] (Y_s^k 는 신뢰도 점수, $\mathcal{L}_s$ 는 패치 이미지에 대한 로짓값, k 는 패치 이미지에서 관찰하려는 클래스, $\xi(\cdot)$ 는 $\mathcal{L}_s$ 와 k 를 입력으로 하여 신뢰도 점수를 계산하는 함수, N 은 클래스의 총 개수, i 는 클래스를 특정하는 인덱스, j 는 로짓을 출력하는 최종 노드를 특정하는 인덱스)

[0074] 다음으로, S1050 단계에서, 장치(100)는 복수의 히트맵 각각이 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 복수의 히트맵이 누적으로 배치되어 차지하는 영역 중 소정의 위치 별(ex. 픽셀 별, 기 설정된 크기의 영역 별)로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성할 수 있다. 설명맵을 생성하는 구체적 예시를 도 6을 참조하여 설명한다.

[0075] 도 6은 일 실시 예에 따라 설명맵을 생성하는 동작에 대한 이해를 돕기 위한 예시도이다.

[0076] 도 6을 참조하면, 장치(100)는 복수의 히트맵 각각(도 6의 좌측 $C_1, C_2, C_3, \dots, C_n$)을 윈도우 사이즈의 크기로 리사이징 하여, 리사이징된 복수의 히트맵(도 6의 우측 $C_1, C_2, C_3, \dots, C_n$) 각각이 타겟 이미지에서 분할된 영

역에 위치하도록 누적하여 배치(도 6의 우측 하단)할 수 있다. 이때, 장치(100)는 복수의 히트맵이 누적으로 배치되어 차지하는 영역 중 어느 제1 위치(ex. 특정 픽셀, 특정 크기의 영역)에 누적된 각 히트맵의 픽셀 정보에 각 히트맵의 신뢰도 점수를 가중치로 가하여 합산 정보를 생성하고, 제1 위치에 누적된 히트맵의 개수를 기초로 합산 정보를 정규화할 수 있다.

[0077] 여기서, 복수의 히트맵 중에는 타겟 이미지의 객체를 온전히 포함하지 않는 히트맵도 존재할 가능성이 있기 때문에, 각 히트맵의 픽셀 정보에 앞서 구한 신뢰도 점수를 가중치로 가하는 것이다. 또한, 복수의 히트맵 각각이 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 히트맵의 수가 다를 수 있기 때문에, 각 위치 별로 누적된 히트맵의 개수를 기초로 합산 정보에 대한 정규화를 수행하는 것이다.

[0078] 이에 따라, S1050 단계에서, 장치(100)는 상술한 제1 위치에서 수행한 동작을 누적으로 배치된 복수의 히트맵이 차지하는 모든 위치에서 수행함으로써 최종 설명맵을 생성할 수 있다. 이때 도 3에서 예시된 설명맵의 좌측은 GradCAM 자체를 사용하여 타겟 이미지로부터 도출된 설명맵이고, 도 3에서 예시된 설명맵의 우측은 GradCAM의 전후 단계에 본 발명의 실시 예를 적용하여 타겟 이미지로부터 도출된 설명맵이다. 도 3에 도시된 설명맵의 예시를 참조하면, 본 발명의 실시 예를 통한 설명맵이 신경망 모델이 참조한 객체의 영역을 보다 온전하게 특정하고 있다는 것을 확인할 수 있다.

[0079] 이러한 비교 예시를 다음 도 7 및 도 8을 통해 더 확인해본다.

[0080] 도 7은 기존의 시각화 방법을 적용하여 설명맵을 생성한 결과와, 기존의 시각화 방법에 본 발명의 실시 예를 적용하여 설명맵을 생성한 결과를 비교한 도면이다.

[0081] 도 7의 비교에서는 시각화 설명 기법 중 GradCAM, FullGrad, RAP, AGF를 사용하여 타겟 이미지(Class: Gold Fish)에 대한 신경망 모델의 시각화를 수행하였다. 도 7의 Original은 GradCAM, FullGrad, RAP, AGF 자체를 사용하여 타겟 이미지로부터 도출된 설명맵이고, 도 7의 Ours는 GradCAM, FullGrad, RAP, AGF의 전후 단계에 본 발명의 실시 예를 적용하여 타겟 이미지로부터 도출된 설명맵이다. 도 7을 참조하면, 본 발명의 실시 예의 설명맵이 기존 기법 자체만을 사용한 설명맵에 비하여 신경망 모델이 참조한 객체의 영역을 보다 온전하게 특정하고 있는 것을 확인할 수 있다.

[0082] 도 8은 ResNet50과 VGG19의 신경망 모델에 대해, 기존의 시각화 방법을 적용하여 설명맵을 생성한 결과와, 기존의 시각화 방법에 본 발명의 실시 예를 적용하여 설명맵을 생성한 결과를 비교한 도면이다.

[0083] 도 8의 비교에서는 시각화 설명 기법 중 GradCAM, RSP, AGF를 사용하여 타겟 이미지(Class: Red Fox, Cheetah, Magpie, Green Snake)에 대한 신경망 모델의 시각화를 수행하였다. 도 8의 타겟 이미지의 우측 상단행은 GradCAM, RSP, AGF 자체를 사용하여 타겟 이미지로부터 도출된 설명맵이고, 도 8의 타겟 이미지의 우측 하단행은 GradCAM, RSP, AGF의 전후 단계에 본 발명의 실시 예를 적용하여 타겟 이미지로부터 도출된 설명맵이다. 도 8을 참조하면, 본 발명의 실시 예의 설명맵이 기존 기법 자체만을 사용한 설명맵에 비하여 신경망 모델이 참조한 객체의 영역을 보다 온전하게 특정하고 있는 것을 확인할 수 있다.

[0084] 도 9는 ResNet50, DenseNet121, InceptionV3의 신경망 모델에 기존의 시각화 방법(GCAM, GCAM++, WGCAM)을 적용한 경우 deletion/insertion game test의 성능과, 기존의 시각화 방법에 본 발명의 실시 예를 적용한 경우((GCAM*, GCAM+*, WGCAM*) deletion/insertion game test성능을 비교한 성능 평가 테이블이다.

[0085] 도 9를 참조하면, 본 발명의 실시 예를 적용한 모든 경우에서 deletion game test의 AUC 점수가 더욱 낮아져 성능이 향상된 것을 확인할 수 있고, insertion game test의 AUC 점수가 더욱 높아져 성능이 향상된 것을 확인할 수 있다.

[0086] 도 10은 다양한 시각화 설명 기법과 본 발명의 실시 예를 적용한 경우, 각 설명맵이 객체를 특정하는 정도에 대한 점수를 비교한 성능 평가 테이블이다.

[0087] 도 10의 비교에서 사용된 본 발명의 실시 예는 GradCAM의 전후 과정에 본 발명의 분할 및 취합 과정을 접목한 실시 예이다. 도 10을 참조하면, 본 발명의 실시 예는 비교 대상 기법에 비하여 월등한 성능 향상을 보이는 것을 확인할 수 있다.

[0088] 상기와 같은 본 발명에 따르면, 타겟 이미지를 분할하여 타겟 이미지의 부분적 정보를 포함하는 패치 이미지를 생성하고, 패치 이미지에 대한 히트맵의 시각화 정보를 취합하여, 세분성과 정확성이 강화된 설명법을 도출함으로써, 신경망 모델이 타겟 이미지의 어떤 영역을 주로 참고하여 특정한 결정을 내리는지 시각화 할 수 있다.

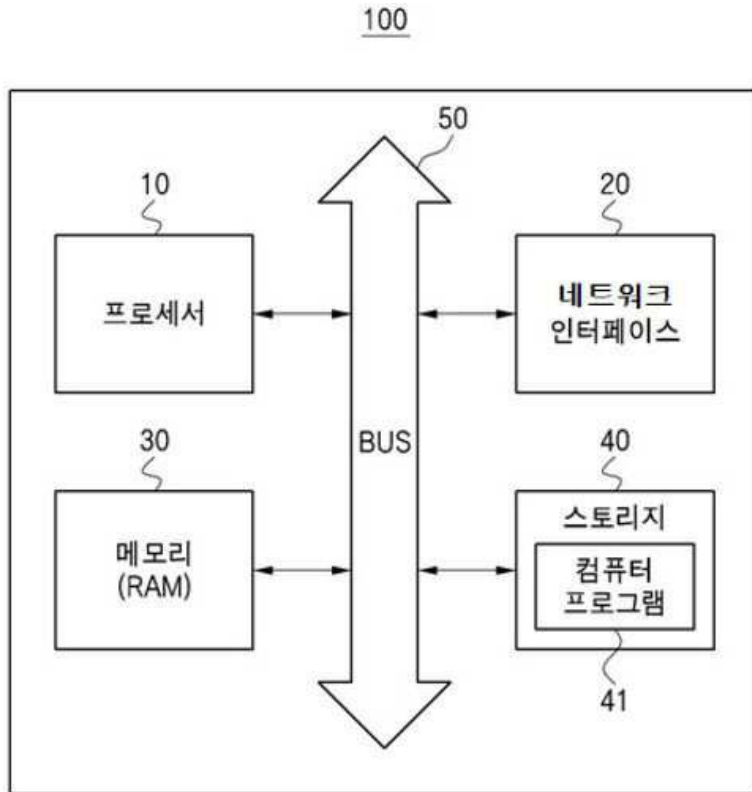
- [0089] 이에 따라, 본 발명은 타겟 이미지에의 어떠한 특징이 신경망 모델의 의사 결정에 주요한 영향을 미치는지 직관적으로 확인할 수 있게 한다. 특히, 의학 분야와 같이 의사 결정의 부담이 큰 분야에서는, 진단의 근거를 제시하여 진단 과정의 정확성과 효율을 향상시킬 수 있으며, 자율 주행과 같이 인명 피해와 직접적으로 연관된 분야에서는 신경망 모델의 의사 결정이 잘못된 특징으로부터 유도되는지 미리 확인하여 수정할 수 있게 한다.
- [0090] 이처럼, 본 발명을 다양한 분야에 활용하는 경우, 사용자는 신경망 모델 및 데이터셋 개선을 위한 방향성을 제 공받아 신경망 모델을 보다 효과적으로 활용할 수 있으며, 나아가 인공지능 신경망 분야의 이해와 발전에 기여할 수 있다.
- [0091] 또한, 상술한 실시 예에 따른 방법은 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램 또는 컴퓨터로 판독 가능한 기록 매체의 형태로도 구현될 수 있다. 일 실시 예에 따른 컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램은 프로세서에 의해 실행되면, 타겟 이미지를 획득하는 동작; 타겟 이미지의 어느 일부 영역을 포함하도록 분할한 복수의 패치 이미지를 생성하는 동작; 상기 복수의 패치 이미지를 소정의 신경망 모델에 입력 시에 상기 모델의 결정을 시각화한 복수의 히트맵을 생성하는 동작; 상기 복수의 히트맵 각각에 대한 신뢰도 점수를 계산하는 동작; 및 상기 복수의 히트맵 각각이 상기 타겟 이미지에서 분할된 영역에 오도록 누적하여 배치할 때, 위치 별로 누적된 각 히트맵의 픽셀 정보에, 누적된 각 히트맵의 신뢰도 점수 및 누적된 히트맵의 개수를 반영하여 설명맵을 생성하는 동작을 포함하는 방법을 프로세서가 수행하도록 하기 위한 명령어를 포함할 수 있다.
- [0092] 또한, 본 문서의 실시 예들의 프로세서는, 저장 매체로부터 저장된 하나 이상의 명령어들 중 적어도 하나의 명령을 호출하고, 그것을 실행할 수 있다. 이것은 기기가 호출된 적어도 하나의 명령어에 따라 적어도 하나의 기능을 수행하도록 운영되는 것을 가능하게 한다. 이러한 하나 이상의 명령어들은 컴파일러에 의해 생성된 코드 또는 인터프리터에 의해 실행될 수 있는 코드를 포함할 수 있다. 프로세서는 범용 프로세서, FPGA(Field Programmable Gate Array), ASIC(Application Specific Integrated Circuit), DSP(Digital Signal Processor) 및/또는 등등 일 수 있다.
- [0093] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

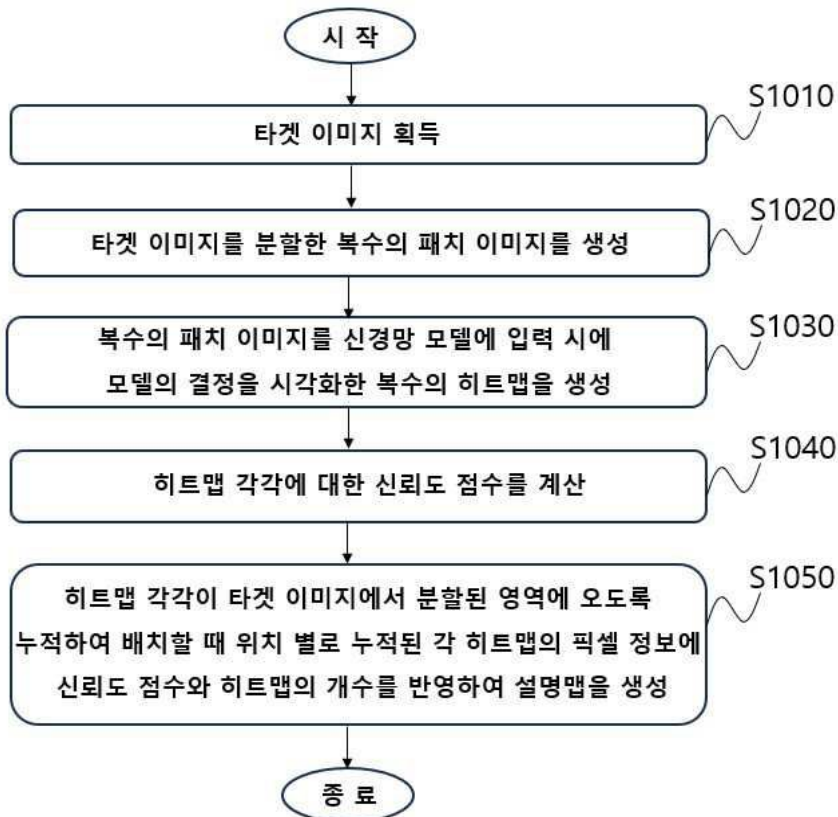
- [0094] 10: 프로세서
20: 네트워크 인터페이스
30: 메모리
40: 스토리지
41: 컴퓨터 프로그램
50: 데이터 버스
100: 시각화 장치

도면

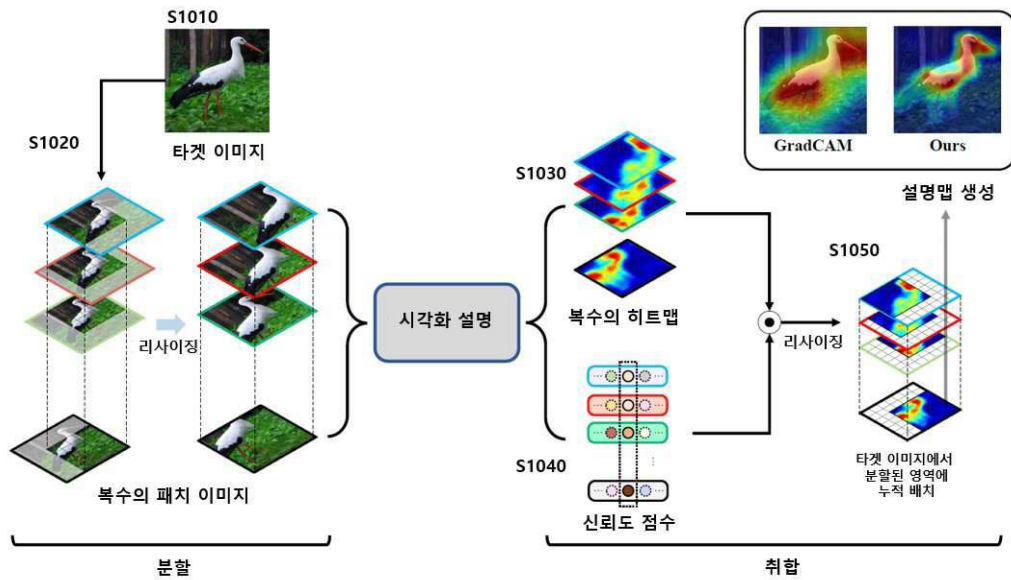
도면1



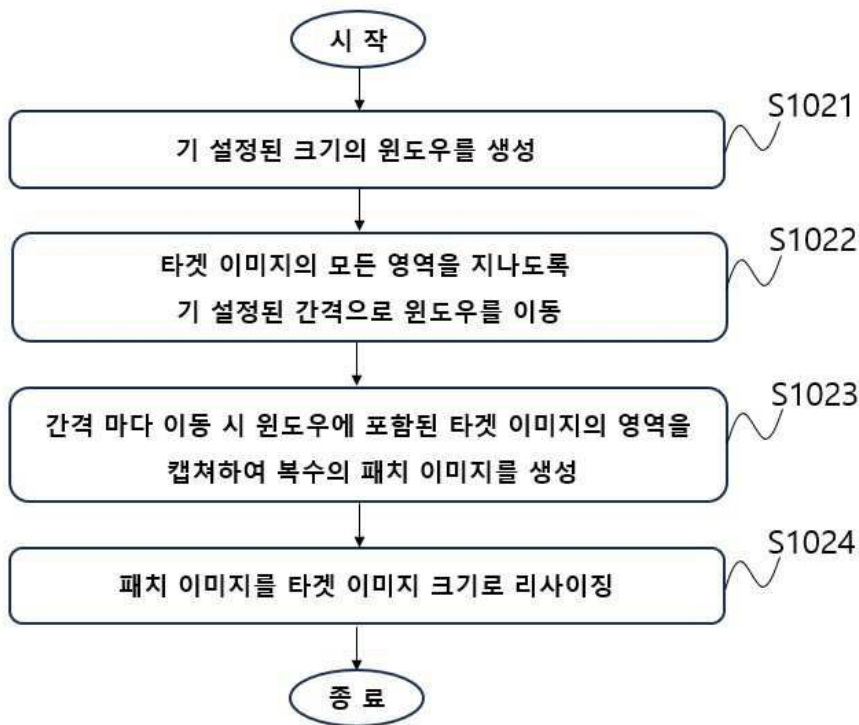
도면2



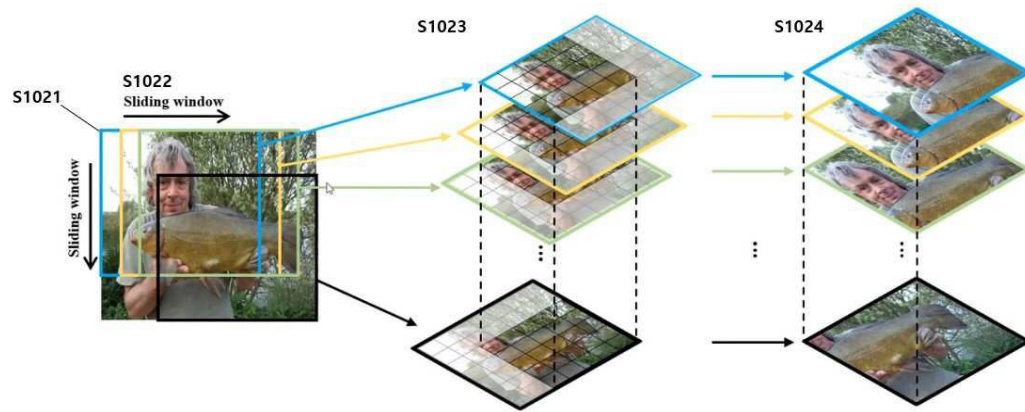
도면3



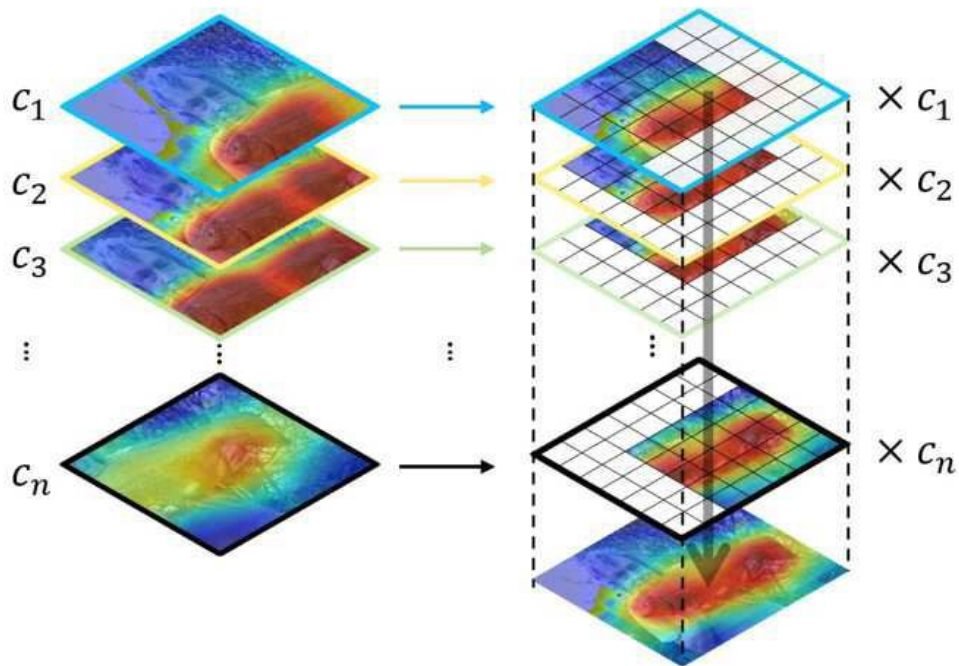
도면4



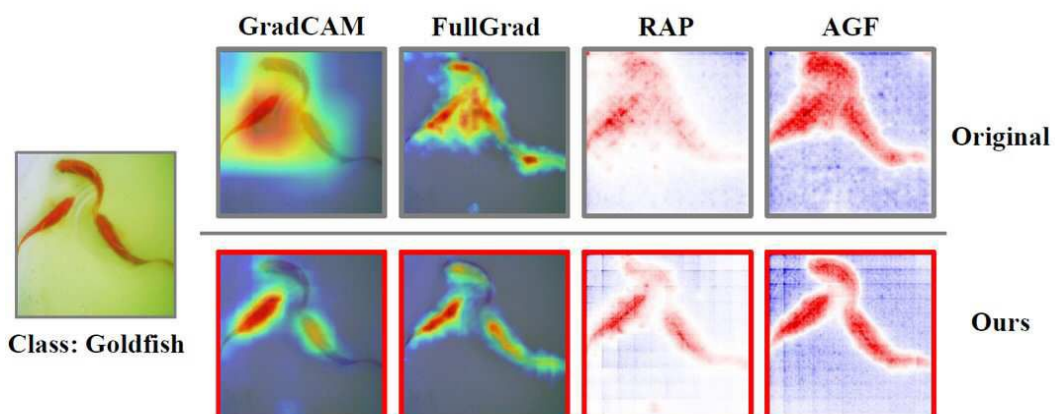
도면5



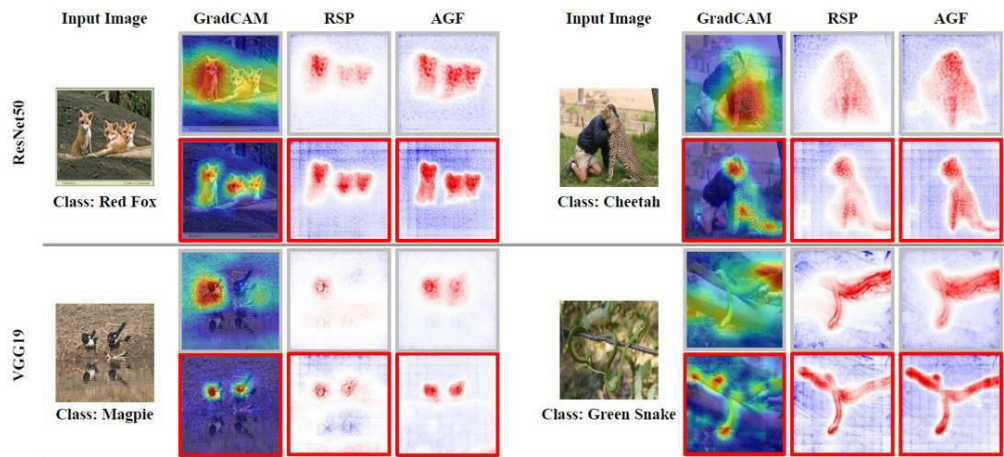
도면6



도면7



도면8



도면9

Methods	ResNet50	DenseNet121	InceptionV3
GCAM	11.3/53.9	10.6/48.1	10.0/52.8
GCAM++	11.6/52.7	10.9/47.2	10.1/52.0
WGCAM	10.1/52.1	10.8/47.7	10.1/52.1
GCAM*	8.6/53.4	8.9/49.0	8.59/53.4
GCAM++*	8.7/52.7	9.2/48.0	8.86/51.9
WGCAM*	9.1/52.8	9.3/48.6	9.08/52.5

도면10

Methods	VOC07 Test		COCO14 Val	
	VGG16	ResNet50	VGG16	ResNet50
Center	69.6/42.4	69.6/42.4	27.8/19.5	27.8/19.5
Gradient	76.3/56.9	72.3/56.8	37.7/31.4	35.0/29.4
DeConv	67.5/44.2	68.6/44.7	30.7/23.0	30.0/21.9
Guid	75.9/53.0	77.2/59.4	39.1/31.4	42.1/35.3
MWP	77.1/56.6	84.4/70.8	39.8/32.8	49.6/43.9
cMWP	79.9/66.5	90.7/82.1	49.7/44.3	58.5/53.6
RISE	86.9/75.1	86.4/78.8	50.8/45.3	54.7/50.0
GradCAM	86.6/74.0	90.4/82.3	54.2/49.0	57.3/52.3
Extremal	88.0/76.1	88.9/78.7	51.5/45.9	56.5/51.5
NormGrad	81.9/64.8	84.6/72.2	-	-
CAMERAS	86.2/76.2	94.2/88.8	55.4/50.7	69.6/66.4
Ours	91.1/82.8	94.2/89.4	61.8/57.6	71.0/67.6