



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0030159
(43) 공개일자 2025년03월05일

(51) 국제특허분류(Int. Cl.) G06N 5/04 (2023.01) G06F 11/36 (2025.01) G06N 3/08 (2023.01) G06N 3/09 (2023.01)	(71) 출원인 고려대학교 산학협력단 서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(52) CPC특허분류 G06N 5/045 (2023.01) G06F 11/362 (2025.01)	(72) 발명자 이성환 서울특별시 강남구 언주로30길 13 (대림아크로빌, 대림아크로텔) B-2902
(21) 출원번호 10-2023-0111188	(74) 대리인 특허법인위솔
(22) 출원일자 2023년08월24일 심사청구일자 2023년08월24일	

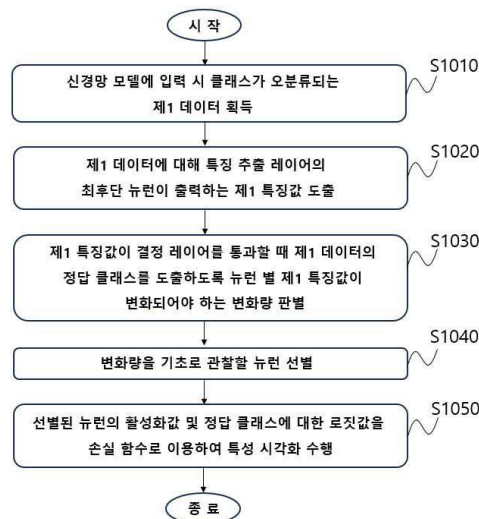
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 클래스 조건 시각화 기반의 이미지 분류 신경망 모델 디버깅 방법 및 이를 위한 장치

(57) 요약

일 실시 예에 따른 디버깅 장치는 특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작; 상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작; 상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작; 상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및 상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용하여 특성 시각화 수행 특성 시각화(Feature Visualization)를 수행하는 동작을 수행할 수 있다.

대표도 - 도2



(52) CPC특허분류

G06N 3/084 (2023.01)

G06N 3/09 (2023.01)

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 디버깅 장치가 수행하는 디버깅 방법에 있어서,

특정 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작;

상기 제1 데이터에 대해 상기 특정 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작;

상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작;

상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및

상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작;

을 포함하는 방법.

청구항 2

제1항에 있어서,

상기 방법은,

상기 특성 시각화를 수행하는 동작 이후에,

상기 선별된 뉴런과 연결된 상기 결정 레이어의 가중치의 영향을 완화하는 방향으로 상기 결정 레이어의 가중치를 재학습시키거나 또는 가중치를 변경하는 동작을 더 포함하는,

방법.

청구항 3

제1항에 있어서,

상기 변화량을 판별하는 동작은

상기 정답 클래스의 로짓을 도출하기 위해 상기 결정 레이어에 입력되어야 하는 제2 특징값을 도출하는 동작; 및

상기 제2 특징값과 상기 제1 특징값의 차이를 기초로 상기 변화량을 판별하는 동작을 포함하는,

방법.

청구항 4

제3항에 있어서,

상기 제2 특징값을 도출하는 동작은

상기 정답 클래스의 로짓을 상기 결정 레이어에 역전파하여 상기 제2 특징값을 도출하는 동작을 포함하는,

방법.

청구항 5

제1항에 있어서,

상기 변화량을 판별하는 동작은

하기 수학식 1의 손실 함수를 최소화하는 방향으로 상기 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작을 포함하는,

[수학식 1]

$$\mathcal{L}_{mis} = \mathcal{L}_{CE}(y, h_t(h_b(x) + \omega)) + \lambda_1 |\omega|_1 + \lambda_2 |\omega|_2$$

($\mathcal{L}_{mis}$ 은 손실 함수, $\mathcal{L}_{CE}$ 는 정답 클래스와 신경망 모델이 예측한 클래스 간의 오차, y 는 정답 클래스, $h_b(x)$ 는 뉴런 x 에 대한 특징 추출기의 제1 특징값, ω 는 제2 특징값이 되기 위한 제1 특징값으로부터의 변화량, h_t 는 결정 레이어의 출력, λ_1 과 λ_2 는 정규화를 위한 하이퍼파라미터)

방법.

청구항 6

제1항에 있어서,

상기 뉴런을 선별하는 동작은

상기 변화량이 큰 순서로 기 설정된 개수의 뉴런을 선별하는 동작을 포함하는,

방법.

청구항 7

제1항에 있어서,

상기 손실 함수는

하기 수학식 2로 구성된,

[수학식 2]

$$\mathcal{L} = -(\alpha_n + \gamma l_c)$$

($\mathcal{L}$ 은 특성 시각화 기법의 손실 함수, α_n 는 최후단 뉴런 n 의 활성화값, l_c 은 클래스 c 에 대한 로짓, γ 은 α_n 과 l_c 의 밸런스를 조절하는 하이퍼파라미터)

방법.

청구항 8

제1항에 있어서,

상기 특성 시각화를 수행하는 동작은

상기 활성화값과 상기 로짓값의 극대화를 조절하는 인자로서, 정답 텍스트와 정답 이미지 간의 유사도를 유지시키는 인자인 하기 수학식 3의 클립 유사도를 더 포함하는 손실 함수를 이용하여 특성 시각화를 수행하는 동작을 포함하는,

[수학식 3]

$$\mathcal{D}_{CLIP} = D_{cos}(E_I(I_{opt}), E_T(T_c))$$

($\mathcal{D}_{CLIP}$ 은 클립 유사도, I_{opt} 은 가우시안 노이즈로부터 생성하고자 하는 이미지의 초기값, T_c 는 클래스

에 대한 텍스트, E_I 는 이미지 인코더, E_T 는 텍스트 인코더, D_{cos} 는 코사인 유사도) 방법.

청구항 9

제8항에 있어서,

상기 손실 함수는

하기 수학적 식 4로 구성된,

[수학적 식 4]

$$\mathcal{L} = -(\mathcal{D}_{CLIP} + \epsilon) \cdot (\alpha_n + \gamma l_c)$$

($\mathcal{L}$ 은 특성 시각화 기법의 손실 함수, $\mathcal{D}_{CLIP}$ 클립 유사도, ϵ 는 도메인 오정렬을 방지하기 위한 하이퍼파라미터, α_n 는 최후단 뉴런 n의 활성화값, l_c 은 클래스 c에 대한 로짓, γ 은 α_n 과 l_c 의 밸런스를 조절하는 하이퍼파라미터)

방법.

청구항 10

명령어를 포함하는 메모리; 및

상기 명령어를 실행하는 프로세서를 포함하고,

상기 프로세서의 동작은,

특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작;

상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작;

상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작;

상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및

상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작;

을 포함하는 장치.

청구항 11

컴퓨터 판독 가능 기록매체에 저장되어 있는 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은, 프로세서에 의해 실행되면,

특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작;

상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작;

상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작;

상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및

상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature

Visualization)를 수행하는 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 클래스 조건 시각화 기반의 이미지 분류 신경망 모델 디버깅 방법 및 이를 위한 장치에 관한 것이다. 보다 자세하게는 특징 분류기가 출력하는 특징값을 관찰하여 오분류에 영향을 미치는 뉴런을 도출하고, 해당 뉴런이 어떠한 부분에 집중하여 이미지를 처리하고 있는지 시각화하여 디버깅을 수행하는 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. 한편, 신경망 모델의 입력과 출력 사이는 블랙박스 형태로, 신경망 모델이 입력으로부터 출력까지의 결정을 어떠한 과정 또는 어떠한 방법으로 처리하는 것인지 직접적으로는 알 수 없다. 그렇기 때문에, 신경망 모델이 특정 선택을 하는 이유를 파악할 수 있다면, 인공지능을 이해하여 보다 개선할 수 있는 가능성이 열리므로, 신경망 모델의 처리 과정을 시각화하여 인간이 이해할 수 있도록 하는 ‘설명 가능한 인공지능’의 연구가 활발하다.

[0003] 한편, 이미지 분류 신경망 모델의 경우, 신경망 모델이 특정 정답으로 분류한 이유를 파악하기 위해, 신경망 모델이 정답을 고를 때 가장 많이 참고한 부분을 시각화하려는 시도가 존재하였다.

[0004] 예를 들어, CAVs(Concept Activation Vectors)는 관찰 대상인 네트워크의 병목 레이어에 대한 선형 분류기들을 훈련시키기 위해, 사용자에게 의해 정의된 데이터로부터 자동적으로 발견된 개념을 활용하는 방식으로, 상위 레벨의 의사 결정 과정 또는 네트워크 내 뉴런을 설명하기 위한 방법으로 사용되고 있다.

[0005] 다만, 일반적으로 이미지 분류 신경망 모델의 디버깅 방법은 신경망 모델의 오분류에 대한 원인을 개선하기 위해 추가적인 데이터셋을 활용해야 한다거나, 신경망 모델의 처리 과정에 대한 시각화 이미지가 인간이 온전히 이해하기가 어렵다는 한계가 있다.

[0006] 이에 따라, 신경망 모델이 이미지의 어떠한 부분을 근거로 오분류를 하는지 인간이 이해할 수 있을 정도로 시각화하여 디버깅 할 수 있는 방법론이 개발된다면, 인공지능 신경망의 이해와 발전에 큰 도움이 될 것으로 예상된다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 등록특허공보 제10-1928208호(2018.12.05)

발명의 내용

해결하려는 과제

[0008] 본 발명이 해결하고자 하는 기술적 과제는 신경망 모델의 오분류에 대한 원인을 파악하기 위한 추가적인 데이터셋의 활용 없이, 특징 분류기가 출력하는 특징값 중 오분류를 일으키는 특징값을 관찰하는 방식을 도입하여, 오분류에 영향을 미치는 뉴런을 도출하고, 도출된 뉴런의 활성화값과 정답 클래스에 대한 로짓값에 집중하여 해당 뉴런이 어떠한 원인에 의해 오분류를 일으키고 있는 지 인간이 이해할 수 있도록 시각화하는 기술을 제공하는 것이다.

[0009] 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

[0010] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 프로세서 및 메모리를 포함하는 디버깅 장치가 수행하는 디버깅 방법은, 특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제 1 데이터를 획득하는 동작; 상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작; 상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작; 상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및 상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작을 포함할 수 있다.

[0011] 일 실시 예에 따르면, 상기 방법은 상기 특성 시각화를 수행하는 동작 이후에, 상기 선별된 뉴런과 연결된 상기 결정 레이어의 가중치의 영향을 완화하는 방향으로 상기 결정 레이어의 가중치를 재학습시키거나 또는 가중치를 변경하는 동작을 더 포함할 수 있다.

[0012] 일 실시 예에 따르면, 상기 변화량을 판별하는 동작은 상기 정답 클래스의 로짓을 도출하기 위해 상기 결정 레이어에 입력되어야 하는 제2 특징값을 도출하는 동작; 및 상기 제2 특징값과 상기 제1 특징값의 차이를 기초로 상기 변화량을 판별하는 동작을 포함할 수 있다.

[0013] 일 실시 예에 따르면, 상기 제2 특징값을 도출하는 동작은 상기 정답 클래스의 로짓을 상기 결정 레이어에 역전파하여 상기 제2 특징값을 도출하는 동작을 포함할 수 있다.

[0014] 일 실시 예에 따르면, 상기 변화량을 판별하는 동작은 하기 수학식 1의 손실 함수를 최소화하는 방향으로 상기 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작을 포함할 수 있다.

[0015] [수학식 1]

$$\mathcal{L}_{mis} = \mathcal{L}_{CE}(y, h_t(h_b(x) + \omega)) + \lambda_1 |\omega|_1 + \lambda_2 |\omega|_2$$

[0017] ($\mathcal{L}_{mis}$ 은 손실 함수, $\mathcal{L}_{CE}$ 는 정답 클래스와 신경망 모델이 예측한 클래스 간의 오차, y 는 정답 클래스, $h_b(x)$ 는 뉴런 x 에 대한 특징 추출기의 제1 특징값, ω 는 제2 특징값이 되기 위한 제1 특징값으로부터의 변화량, h_t 는 결정 레이어의 출력, λ_1 과 λ_2 는 정규화를 위한 하이퍼파라미터)

[0018] 일 실시 예에 따르면, 상기 뉴런을 선별하는 동작은 상기 변화량이 큰 순서로 기 설정된 개수의 뉴런을 선별하는 동작을 포함할 수 있다.

[0019] 일 실시 예에 따르면, 상기 손실 함수는 하기 수학식 2로 구성될 수 있다.

[0020] [수학식 2]

$$\mathcal{L} = -(\alpha_n + \gamma l_c)$$

[0022] ($\mathcal{L}$ 은 특성 시각화 기법의 손실 함수, α_n 는 최후단 뉴런 n 의 활성화값, l_c 은 클래스 c 에 대한 로짓, γ 은 α_n 과 l_c 의 밸런스를 조절하는 하이퍼파라미터)

[0023] 일 실시 예에 따르면, 상기 특성 시각화를 수행하는 동작은 상기 활성화값과 상기 로짓값의 극대화를 조절하는 인자로서, 정답 텍스트와 정답 이미지 간의 유사도를 유지시키는 인자인 하기 수학식 3의 클립 유사도를 더 포함하는 손실 함수를 이용하여 특성 시각화를 수행하는 동작을 포함할 수 있다.

[0024] [수학식 3]

$$\mathcal{D}_{CLIP} = D_{cos}(E_I(I_{opt}), E_T(T_c))$$

[0025]

[0026] ($\mathcal{D}_{\text{CLIP}}$ 은 클립 유사도, I_{opt} 은 가우시안 노이즈로부터 생성하고자 하는 이미지의 초기값, T_c 는 클래스에 대한 텍스트, E_I 는 이미지 인코더, E_T 는 텍스트 인코더, D_{cos} 는 코사인 유사도)

[0027] 일 실시 예에 따르면, 상기 손실 함수는 하기 수학적 식 4로 구성될 수 있다.

[0028] [수학적 식 4]

$$\mathcal{L} = -(\mathcal{D}_{\text{CLIP}} + \epsilon) \cdot (\alpha_n + \gamma l_c)$$

[0030] ($\mathcal{L}$ 은 특성 시각화 기법의 손실 함수, $\mathcal{D}_{\text{CLIP}}$ 클립 유사도, ϵ 는 도메인 오정렬을 방지하기 위한 하이퍼파라미터, α_n 는 최후단 뉴런 n의 활성화값, l_c 은 클래스 c에 대한 로짓, γ 은 α_n 과 l_c 의 밸런스를 조절하는 하이퍼파라미터)

[0031] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 디버깅 장치는 명령어를 포함하는 메모리; 및 상기 명령어를 실행하는 프로세서를 포함하고, 상기 프로세서의 동작은, 특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작; 상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작; 상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작; 상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및 상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작을 포함할 수 있다.

[0032] 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, 특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작; 상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작; 상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작; 상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및 상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작을 포함하는 방법을 상기 프로세서가 수행하도록 하기 위한 명령어를 포함할 수 있다.

발명의 효과

[0033] 상기와 같은 본 발명에 따르면, 오분류가 발생하는 제1 데이터에 대해 신경망 모델의 특징 분류기가 출력하는 제1 특징값과 정답 클래스를 도출하기 위한 제2 특징값의 차이에 기반하여, 오분류에 주된 영향을 미치는 뉴런을 도출할 수 있으며, 해당 뉴런의 활성화값과 정답 클래스에 대한 로짓값에 집중하여 해당 뉴런이 어떠한 원인에 의해 오분류를 일으키고 있는 지 인간이 온전히 이해할 수 있는 시각화 기술을 제공할 수 있다.

[0034] 이에 따라, 본 발명은 추가적인 데이터 요구 또는 네트워크 수정 없이 신경망 모델의 오류를 설명할 수 있고, 특정 클래스의 거짓된 상관 관계를 완화시키도록 결정 레이어를 학습시키는 간단한 방식으로 신경망 모델의 성능을 대폭 향상시킬 수 있다.

[0035] 이처럼, 본 발명은 간단한 편집 또는 재학습으로 신경망 모델의 오류를 해결할 수 있기에, 실용적인 디버깅 프레임 워크로 채택될 수 있으며, 결과와 관계된 뉴런의 클래스 조건 시각화를 통해 실무자에게 신경망 모델의 오류에 대한 주된 원인의 분석할 수 있게 한다. 이에 따라, 실무자는 본 발명을 활용하여 모델 및 데이터셋 개선을 위한 방향성을 제공받아 신경망 모델을 보다 효과적으로 활용할 수 있으며, 나아가 신경망 모델의 이해와 발전에 기여할 수 있다.

[0036] 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

- [0037] 도 1은 일 실시 예에 따른 디버깅 장치가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- 도 2는 일 실시 예에 따른 디버깅 장치가 수행하는 디버깅 방법의 대표적인 단계를 나타낸 순서도이다.
- 도 3은 도 2의 실시 예에 따른 디버깅 방법의 각 동작에 대한 이해를 돕기 위해 예시한 개념도이다.
- 도 4는 일 실시 예에 따라 특정 클래스의 오분류를 발생시키는 특징값 변화량이 큰 5개의 뉴런을 선별한 결과를 예시한 도면이다.
- 도 5는 신경망 모델에 포함된 동일한 뉴런 정보를 기준으로, 기존 FV 시각화 기법을 수행한 결과, 실제 이미지, 본 문서의 실시 예에 따라 시각화를 수행한 결과를 비교한 도면이다.
- 도 6은 일 실시 예에 따라 특성 시각화 수행 시, 다양한 손실 함수를 사용하여 특성 시각화를 수행한 결과를 비교한 도면이다.
- 도 7은 도 2의 실시 예를 수행한 이후, 일 실시 예에 따른 디버깅 장치가 추가적으로 수행할 수 있는 디버깅 방법의 단계를 나타낸 순서도이다.
- 도 8은 도 6의 실시 예에 따른 디버깅 방법의 각 동작에 대한 이해를 돕기 위해 예시한 개념도이다.
- 도 9는 특정 클래스에 대한 오분류를 특정한 이후, 단순히 오차를 최소화하는 방향으로 손실 함수를 이용하여 재학습하는 방법과, 본 발명의 일 실시 예에 따라 결정 레이어를 재학습한 경우의 성능 향상을 비교한 성능 평가 테이블이다.
- 도 10은 하나의 클래스를 세분화하여 분류하는 신경망 모델의 기본 성능과, Class weighting, Focal loss, 본 발명의 실시 예를 적용한 경우, 각 기법에 따른 성능 향상을 나타낸 성능 평가 테이블이다.

발명을 실시하기 위한 구체적인 내용

- [0038] * * * * * 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.
- [0039] 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.
- [0040] 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.
- [0041] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.
- [0042] 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.
- [0043] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.
- [0044] 도 1은 일 실시 예에 따른 디버깅 장치(100)(이하, '장치(100)'로 지칭)가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- [0045] 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.
- [0046] 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.

- [0047] 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.
- [0048] 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.
- [0049] 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.
- [0050] 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.
- [0051] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해, 특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작; 상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작; 상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작; 상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및 상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작을 실행할 수 있다.
- [0052] 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.
- [0053] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의 명령 및/또는 정보의 이동 경로가 된다.
- [0054] 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.
- [0055] 이하, 일 실시 예에 따른 장치(100)가 수행하는 방법에 대하여 도 2 내지 도 10을 참조하여 설명하도록 한다.
- [0056] 도 2는 일 실시 예에 따라 장치(100)가 수행하는 디버깅 방법의 대표적인 단계를 나타낸 순서도이고, 도 3은 도 2의 실시 예에 따른 디버깅 방법의 각 단계에 대한 이해를 돕기 위해 예시한 개념도이다.
- [0057] 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.
- [0058] 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.
- [0059] 도 2 및 도 3에 개시된 각 동작의 순서는 이해의 편의를 위해 배치된 순서일 뿐, 이러한 순서가 시계열적인 순

서로 한정되는 것이 아니며, 설계자의 선택에 따라 순서가 다르게 변경되어 동작될 수 있다.

- [0060] 도 2 및 도 3을 함께 참조하면, S1010 단계에서, 장치(100)는 입력되는 이미지의 클래스를 분류하도록 학습된 신경망 모델을 획득하여, 해당 신경망 모델에 데이터를 입력하는 경우, 해당 신경망 모델이 클래스를 오분류하여 출력하게 되는 데이터인 제1 데이터를 획득할 수 있다.
- [0061] 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.
- [0062] 신경망 모델은 입력된 데이터의 특징값(Feature)을 추출하는 특징 추출기(Feature Extractor) 및 특징 추출기가 추출한 특징값을 기초로 입력된 데이터의 클래스를 결정하는 결정 레이어(Decision Layer)로 구성될 수 있다.
- [0063] 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.
- [0064] 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.
- [0065] 본 문서의 실시 예는 특징 분류기가 출력하는 특징값 중 결정 레이어의 오분류를 일으키는 제1 데이터의 특징값을 판별하고, 특징 분류기를 구성하는 뉴런 중 제1 데이터로부터 정답 클래스가 도출하도록 변화시켜야 하는 뉴런을 선별하기 위하여 다음의 동작을 수행한다.
- [0066] 다음으로, S1020 단계에서, 장치(100)는 제1 데이터에 대해 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출할 수 있다. 예를 들어, 특징 추출기는 합성곱층과(Convolution layer)와 풀링층(Pooling layer)로 구성될 수 있고, 최후단 뉴런은 특징 추출기를 구성하는 레이어 중 가장 마지막(최후단)에 위치한 레이어에 포함된 뉴런을 포함할 수 있다.
- [0067] 이에, S1030 단계에서, 장치(100)는 제1 특징값이 결정 레이어를 통과할 때 제1 데이터의 정답 클래스를 도출하도록 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별할 수 있다. 예를 들어, 장치(100)는 정답 클래스의 로짓(logit)을 도출하기 위해 결정 레이어에 입력되어야 하는 최후단 뉴런 별 제2 특징값을 도출하여, 제2 특징값과 제1 특징값의 차이를 기초로 뉴런 별 특징값 변화량을 판별할 수 있다.
- [0068] 일 예로, 장치(100)는 정답 클래스의 로짓을 결정 레이어에 역전파하여, 결정 레이어에서 정답 클래스를 도출하도록 하는 제2 특징값을 도출하여, 제2 특징값과 제1 특징값의 차이를 기초로 최후단 뉴런 별 특징값 변화량을 판별할 수 있다. 이 경우, 장치(100)는 각 뉴런이 제2 특징값을 출력하도록, 제1 특징값으로부터 제2 특징값이 되기 위한 변화량 ω 를, 하기 수학적 1의 손실 함수를 최소화함으로써 판별할 수 있다.

수학적 1

[0069]
$$\mathcal{L}_{mis} = \mathcal{L}_{CE}(y, h_t(h_b(x) + \omega)) + \lambda_1 |\omega|_1 + \lambda_2 |\omega|_2$$

[0070] ($\mathcal{L}_{mis}$ 은 손실 함수, $\mathcal{L}_{CE}$ 는 정답 클래스와 신경망 모델이 예측한 클래스 간의 오차, y 는 정답 클래스, $h_b(x)$ 는 뉴런 x에 대한 특징 추출기의 제1 특징값, ω 는 제2 특징값이 되기 위한 제1 특징값으로부터의 변화량, h_t 는 결정 레이어의 출력, λ_1 (ex. 0.1)과 λ_2 (ex. 0.01)는 정규화를 위한 하이퍼파라미터)

[0071] 다음으로, S1040 단계에서, 장치(100)는 S1030 단계에서 도출된 각 뉴런 별 특징값 변화량을 기초로 관찰할 뉴런을 선별할 수 있다. 이때 특징값 변화량은 제1 특징값에 비하여 수치가 더 증가하여야 하는 (+) 방향의 변화량과, 제1 특징값에 비하여 수치가 더 감소하여야 하는 (-) 방향의 변화량을 포함할 수 있다. 예를 들어, 장치(100)는 (+) 방향의 변화량이 큰 순위(ex. 도 3의 #510, #481, #56)와, (-) 방향의 변화량이 큰 순위의 뉴런

(ex. 도 3의 #452, #245)를 각각 구분하여 뉴런을 선별할 수 있다.

[0072] 또한, 장치(100)는 뉴런 별 특징값 변화량이 큰 순서로 뉴런의 순위를 매기고, 기 설정된 개수만큼 순위가 큰 순서로 뉴런을 선별할 수 있다. 이러한 예시를 도 4와 함께 설명한다.

[0073] 도 4는 일 실시 예에 따라 특정 클래스의 오분류를 발생시키는 특징값 변화량이 큰 5개의 뉴런을 선별한 결과를 예시한 도면이다. 도 4의 Class는 정답 클래스, correlation은 정답 객체와 연관되어 나타나는 타 객체를 의미하며, 도 4에는 5개의 뉴런 번호 별 특징값에 의한 correlation의 영향력이 도시되어 있다.

[0074] 도 4를 참조하면, 특징값 변화량이 크다는 것은 정답 클래스를 도출하기 위해 해당 뉴런이 출력하는 특징값이 개선되어야 하는 정도가 크다는 것을 의미하며, 순위를 기초로 뉴런을 선별하는 이유는 특징값 개선 정도가 큰 뉴런을 집중적으로 관찰하기 위함이다. 일반적으로, 신경망 모델은 수백만 개의 뉴런을 포함할 수 있고, 각 뉴런의 특성을 관찰하는 것은 매우 오랜 시간이 걸리기 때문에, 도 4와 같이 특징값이 개선되어야 하는 정도가 큰 순서로 기 설정된 개수(ex. 5개)의 뉴런을 선별할 수 있다.

[0075] 이에, S1050 단계에서, 장치(100)는 선별된 뉴런의 활성화값 및 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행할 수 있다.

[0076] 여기서, 특성 시각화란 신경망 모델의 뉴런에 대한 학습 특성을 시각화하는 기법이다. 예를 들어, 장치(100)는 특성 시각화 기법을 이용하여, 가우시안 노이즈로부터 생성하고자 하는 이미지 크기의 랜덤한 텐서 I_{opt} 를 생성하고, 관찰하고자 하는 뉴런의 활성화 α_n 를 높이도록 설계된 손실 함수를 기초로 텐서 I_{opt} 를 업데이트함으로써, 해당 뉴런을 극도로 활성화시킨 이미지를 생성하여 해당 뉴런의 특징을 시각화할 수 있다.

[0077] 일 예로, S1050 단계에서, 장치(100)는 특성 시각화 기법을 적용하여 이미지를 생성할 때, 관찰할 뉴런의 활성화값 및 클래스 조건을 극대화하도록 하는 방향으로, 하기 수학식 2의 손실 함수를 이용해 텐서 이미지 I_{opt} 를 업데이트하여 뉴런의 시각화를 수행할 수 있다.

수학식 2

$$\mathcal{L} = -(\alpha_n + \gamma l_c)$$

[0078]

[0079] ($\mathcal{L}$ 은 특성 시각화 기법의 손실 함수, α_n 는 최후단 뉴런 n의 활성화값, l_c 은 클래스 c에 대한 로짓, γ 은 α_n 과 l_c 의 밸런스를 조절하는 하이퍼파라미터)

[0080] 일 예로, S1050 단계에서, 장치(100)는 특성 시각화 기법을 적용하여 이미지를 생성할 때, 관찰할 뉴런의 활성화값과 클래스 조건의 과도한 극대화를 방지하는 인자로서, 정답 텍스트와 정답 이미지 간의 유사도를 유지시키는 인자인 하기 수학식 3의 클립 유사도를 더 포함하는 하기 수학식 4의 손실 함수를 이용하여 텐서 이미지 I_{opt} 를 업데이트함으로써 뉴런의 시각화를 수행할 수 있다.

수학식 3

$$\mathcal{D}_{CLIP} = D_{cos}(E_I(I_{opt}), E_T(T_c))$$

[0081]

[0082] ($\mathcal{D}_{CLIP}$ 은 클립 유사도, I_{opt} 은 가우시안 노이즈로부터 생성하고자 하는 이미지의 초기값, T_c 는 클래스에 대한 정답 텍스트, E_I 는 이미지 인코더, E_T 는 텍스트 인코더, D_{cos} 은 코사인 유사도)

수학식 4

$$\mathcal{L} = -(\mathcal{D}_{\text{CLIP}} + \epsilon) \cdot (\alpha_n + \gamma l_c)$$

[0083]

[0084]

($\mathcal{L}$ 은 특성 시각화 기법의 손실 함수, $\mathcal{D}_{\text{CLIP}}$ 클립 유사도, ϵ 는 도메인 오정렬을 방지하기 위한 하이퍼파라미터, α_n 는 최후단 뉴런 n의 활성화값, l_c 은 클래스 c에 대한 로짓, γ 은 α_n 과 l_c 의 밸런스를 조절하는 하이퍼파라미터)

[0085]

도 5는 신경망 모델에 포함된 동일한 뉴런 정보를 기준으로, 기존의 FV 시각화 기법을 수행한 결과(도 4의 FV 열), 실제 이미지(도 4의 Real Image 열), 본 문서의 실시 예에 따라 시각화를 수행한 결과(도 4의 CLIP-Illusion 열)를 비교한 도면이다.

[0086]

도 5를 참조하면, 특정 클래스를 판별하는 뉴런의 특성을 시각화 할 때, 기존 시각화 기법에 의해 생성된 이미지에 비하여, 본 문서의 실시 예를 통해 생성한 이미지가 인간이 인지 가능한 형태로 시각화된다는 것을 확인할 수 있다.

[0087]

도 6은 일 실시 예에 따라 특성 시각화 수행 시, 다양한 손실 함수를 사용하여 특성 시각화를 수행한 결과를 비교한 도면이다.

[0088]

도 6에서 첫번째 행은 기존 FV 기법의 손실 함수, 두번째 행은 상술한 수학식 2의 손실 함수, 세번째 행은 상술한 수학식 4의 손실 함수를 의미하며, 도 6의 각 열 최상단은 신경망 모델과 뉴런 정보를 의미하며, 도 6의 각 열 최하단은 정답 클래스를 의미한다. 도 6을 참조하면, 수학식 2와 수학식 4의 손실 함수를 이용하여 뉴런을 시각화한 경우가, 기존 FV 기법에 비해 인간이 인지 가능한 형태로 시각화된다는 것을 확인할 수 있다.

[0089]

도 7은 도 2의 실시 예를 수행한 이후, 일 실시 예에 따른 장치(100)가 추가적으로 수행할 수 있는 디버깅 방법의 단계를 나타낸 순서도이고, 도 8은 도 7의 실시 예에 따른 디버깅 방법의 각 동작에 대한 이해를 돕기 위해 예시한 개념도이다.

[0090]

도 7 및 도 8에 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.

[0091]

도 7 및 도 8에 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.

[0092]

도 7 및 도 8에 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.

[0093]

도 7 및 도 8을 함께 참조하면, S2010 단계에서, 장치(100)는 클래스의 오분류에 영향이 큰 뉴런을 선별할 수 있다. 예를 들어, 장치(100)는 정답 클래스(ex. 개)를 분류하도록 학습된 신경망 모델에 포함된 뉴런을 시각화된 이미지에서 정답 객체와 다른 타 객체(ex. 물)를 포함하는 이미지를 출력하는 뉴런을 선별하거나, 특징별 변화량이 큰 뉴런을 선별할 수 있다.

[0094]

다음으로, S2020 단계에서, 장치(100)는 선별된 뉴런과 연결된 결정 레이어의 가중치의 영향을 완화하는 방향으로 결정 레이어 가중치를 재학습 시키거나, 결정 레이어의 가중치를 변경할 수 있다.

[0095]

도 9는 특정 클래스에 대한 오분류를 특정한 이후, 단순히 오차를 최소화하는 방향으로 손실 함수를 이용하여 재학습하는 방법(CON)과, 본 발명의 일 실시 예에 따라 결정 레이어를 재학습한 경우(Ours)의 성능 향상을 비교한 성능 평가 테이블이다.

[0096]

도 9를 참조하면, Δ 는 본래 신경망 모델의 성능에 대한 디버깅 이후 개선된 성능의 변화량을 나타낸다. 또한, Acc가 높을수록 정규화 효과가 좋고 모델의 오차가 작다는 것을 의미하고, Prec@3 및 Concept score의 값이 낮을수록 목표 스푸리어스 상관관계가 더 완화되었음을 의미한다. 도 9에 따르면 본 발명의 실시 예는 비교 대상 기법에 비하여, Acc, Prec@3, Concept score 모두에서 월등한 향상을 보이는 것을 확인할 수 있다.

- [0097] 도 10은 하나의 클래스를 세분화하여 분류하는 신경망 모델의 기본 성능(Original)과, Class weighting, Focal loss, 본 발명의 실시 예(Ours)를 적용한 경우, 각 기법에 따른 성능 향상을 나타낸 성능 평가 테이블이다.
- [0098] 도 10의 Class Acc는 오분류된 클래스에 대한 성능 향상도, Avg Acc는 평균적인 클래스 분류 성공도, Min Acc는 클래스 중 가장 낮은 성공률로 분류되는 클래스에 대한 성능의 향상도를 의미한다. 도 10을 참조하면, 본 발명의 실시 예(Ours)를 적용하여 디버깅한 경우, 평균적인 클래스 분류 성공도 Avg Acc를 유지하면서도, 오분류된 클래스에 대한 성능 향상도 Class Acc를 대폭 향상시킬 수 있으면서, 동시에 클래스 중 가장 낮은 성공률로 분류되는 클래스에 대한 성능 Min Acc를 대폭 향상시킬 수 있음을 확인할 수 있다.
- [0099] 상기와 같은 본 발명에 따르면, 오분류가 발생하는 제1 데이터에 대해 신경망 모델의 특징 분류기가 출력하는 제1 특징값과 정답 클래스를 도출하기 위한 제2 특징값의 차이에 기반하여, 오분류에 주된 영향을 미치는 뉴런을 도출할 수 있으며, 해당 뉴런의 활성화값과 정답 클래스에 대한 로짓값에 집중하여 해당 뉴런이 어떠한 원인에 의해 오분류를 일으키고 있는 지 인간이 온전히 이해할 수 있는 시각화 기술을 제공할 수 있다.
- [0100] 이에 따라, 본 발명은 추가적인 데이터 요구 또는 네트워크 수정 없이 신경망 모델의 오류를 설명할 수 있고, 특정 클래스의 거짓된 상관 관계를 완화시키도록 결정 레이어를 학습시키는 간단한 방식으로 신경망 모델의 성능을 대폭 향상시킬 수 있다.
- [0101] 이처럼, 본 발명은 간단한 편집 또는 재학습으로 신경망 모델의 오류를 해결할 수 있기에, 실용적인 디버깅 프레임 워크로 채택될 수 있으며, 결과와 관계된 뉴런의 클래스 조건 시각화를 통해 실무자에게 신경망 모델의 오류에 대한 주된 원인의 분석할 수 있게 한다. 이에 따라, 실무자는 본 발명을 활용하여 모델 및 데이터셋 개선을 위한 방향성을 제공받아 신경망 모델을 보다 효과적으로 활용할 수 있으며, 나아가 신경망 모델의 이해와 발전에 기여할 수 있다.
- [0102] 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, 특징 추출기 및 결정 레이어로 구성된 신경망 모델에 입력 시 클래스가 오분류되는 제1 데이터를 획득하는 동작; 상기 제1 데이터에 대해 상기 특징 추출기의 최후단 뉴런이 출력하는 제1 특징값을 도출하는 동작; 상기 제1 특징값이 상기 결정 레이어를 통과할 때 상기 제1 데이터의 정답 클래스를 도출하도록 상기 뉴런 별 제1 특징값이 변화되어야 하는 변화량을 판별하는 동작; 상기 변화량을 기초로 관찰할 뉴런을 선별하는 동작; 및 상기 선별된 뉴런의 활성화값 및 상기 정답 클래스에 대한 로짓값을 손실 함수로 이용한 특성 시각화(Feature Visualization)를 수행하는 동작을 포함하는 방법을 프로세서가 수행하도록 하기 위한 명령어를 포함할 수 있다.
- [0103] 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.
- [0104] 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

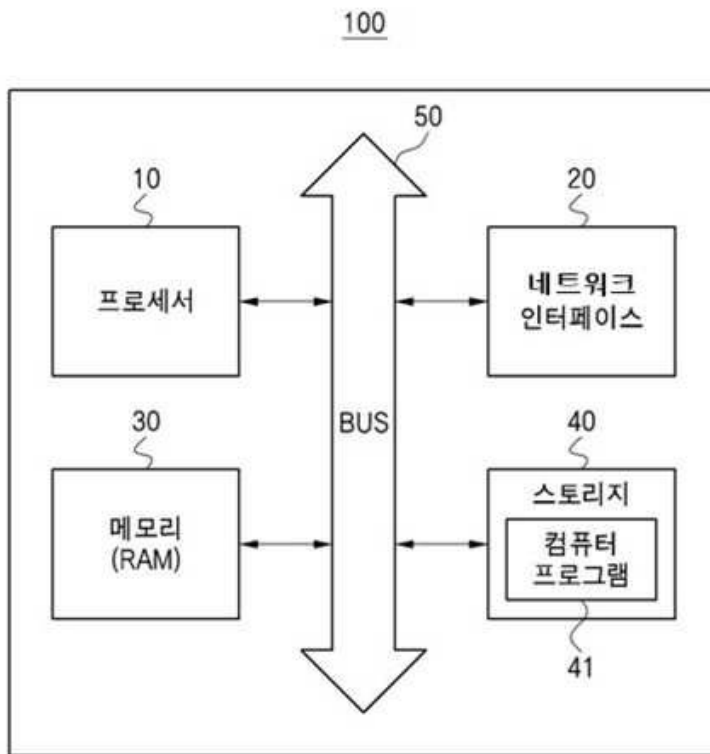
- [0105] 10: 프로세서
20: 네트워크 인터페이스
30: 메모리
40: 스토리지
41: 컴퓨터 프로그램

50: 데이터 버스

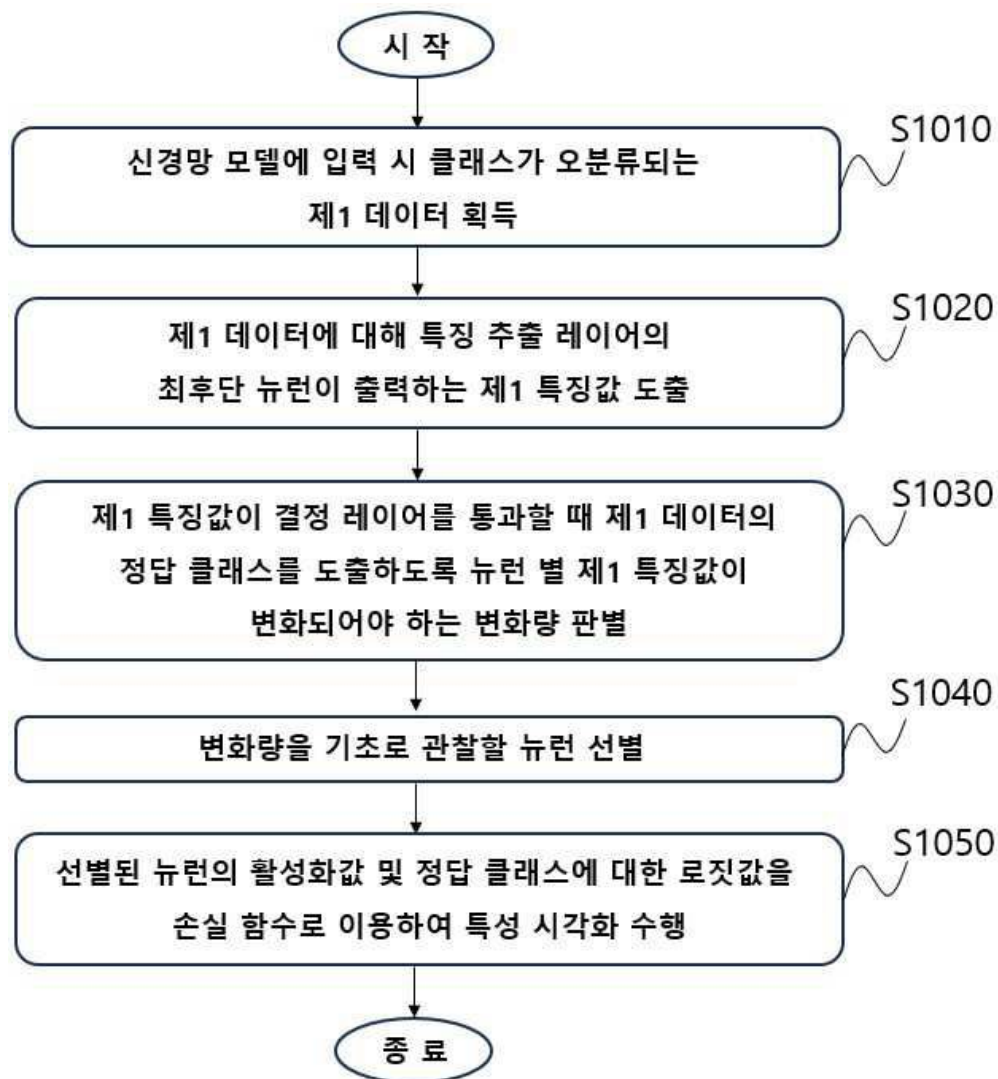
100: 디버깅 장치

도면

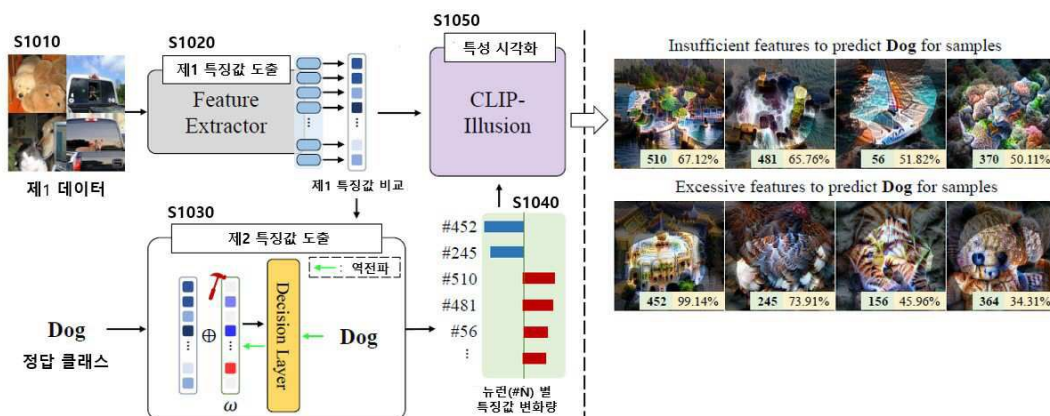
도면1



도면2



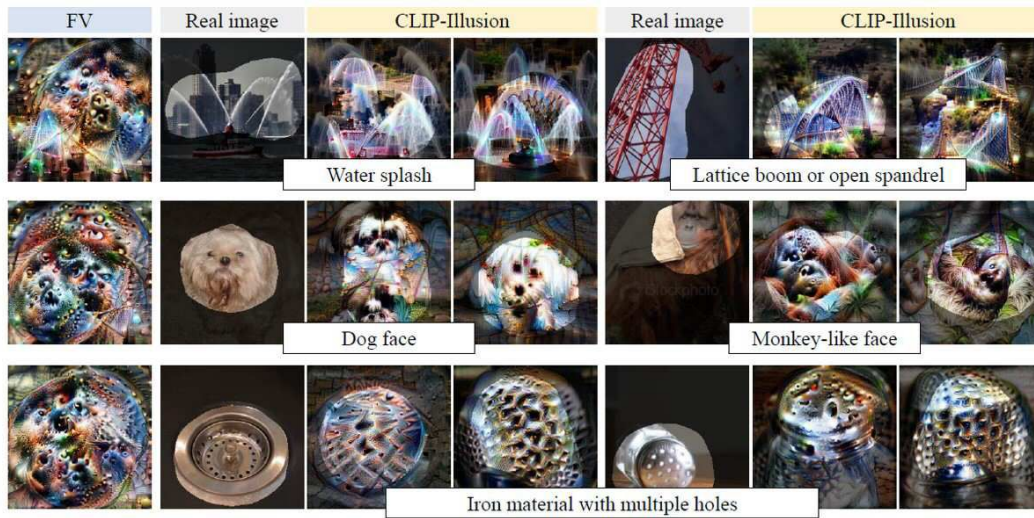
도면3



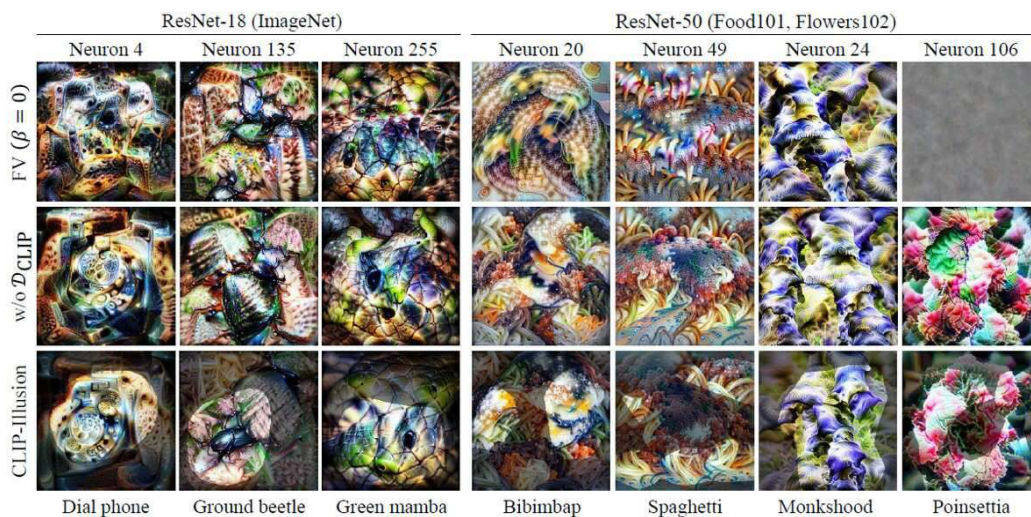
도면4

Class (+correlation)	Chosen neurons (top-5 rank)
Dog (+chair)	244 (56.25%), 36 (36.26%), 47 (31.33%), 430 (21.05%), 194 (4.11%)
Cat (+book)	364 (54.09%), 381 (34.75%), 484 (21.74%), 47 (21.38%)
Bird (+sand)	104 (97.69%), 182 (73.45%), 56 (35.03%), 304 (13.12%), 435 (11.26%)
Dog (+water)	510 (67.12%), 481 (65.76%), 56 (51.82%), 370 (50.10%), 133 (27.19%)
Cat (+keyboard)	222 (79.51%), 121 (78.84%), 201 (42.85%), 373 (28.70%)

도면5



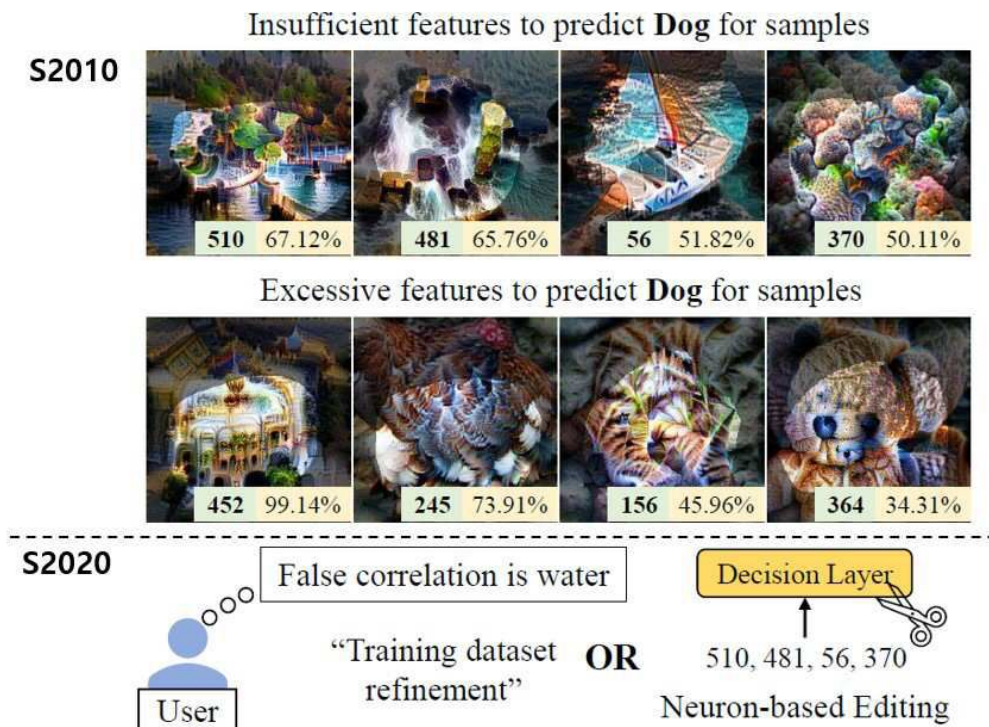
도면6



도면7



도면8



도면9

Class (+correlation) Method	% Acc	% Prec@3 -	Concept score	% Δ Acc ↑ CON Ours	% Δ Prec@3 ↓ CON Ours	Δ Concept score ↓ CON Ours
Dog (+chair)	49.35	81.7	0.492	-2.74 +12.37	-74.4 -78.0	-0.052 -0.162
Cat (+book)	67.61	93.5	0.526	-4.09 +2.65	-41.4 -53.0	-0.1 -0.113
Bird (+sand)	58.51	94.4	0.589	-6.53 +19.27	-4.4 -5.1	-0.1 -0.108
Dog (+water)	39.22	85.0	0.642	+4.65 +3.22	-6.9 -43.3	-0.077 -0.108
Cat (+keyboard)	54.53	38.3	0.432	+7.66 +23.29	-11.4 -30.3	-0.076 -0.098

도면10

Datasets		Flowers102			Food101		
Methods	Class Acc	Avg Acc	Min Class Acc	Class Acc	Avg Acc	Min Class Acc	
Original	33.33%	84.57%	33.33% (3)	43.2%	82.65%	43.2% (93)	
Class weighting	33.33%	84.64%	33.33% (3)	58.8%	82.85%	48.4% (0)	
Focal loss	33.33%	84.05%	33.33% (3)	46.0%	82.42%	46.0% (93)	
Ours	61.11%	84.05%	46.36% (42)	61.2%	81.54%	50.8% (77)	