



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0140681
(43) 공개일자 2023년10월10일

(51) 국제특허분류(Int. Cl.)
G06N 5/04 (2023.01) G06N 3/04 (2023.01)
G06N 5/00 (2023.01)
(52) CPC특허분류
G06N 5/045 (2023.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2022-0039250
(22) 출원일자 2022년03월30일
심사청구일자 2022년03월30일

(71) 출원인
계명대학교 산학협력단
대구광역시 달서구 달구벌대로 1095, 계명대학교
산학협력관 201호(신당동)
(72) 발명자
고병철
대구광역시 수성구 범어천로 107 범어 STX KAN,
101동 1002호
김상원
대구광역시 달서구 달구벌대로 1095 계명대학교
(74) 대리인
김건우

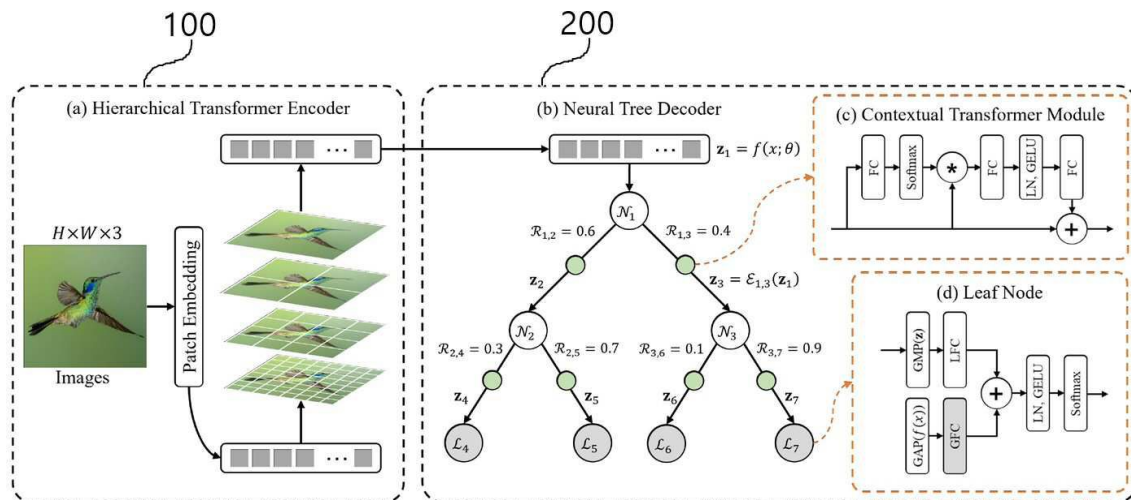
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법

(57) 요약

본 발명은 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기에 관한 것으로서, 보다 구체적으로는 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기(ViT-NeT)로서, 입력 이미지에서 주의 맵(attention map) 특징을 추출하는 계층적 변환기 인코더(hierarchical transformer encoder); 및 리프 노드를 포함하는 복수의 노드 및 상위(뒷면에 계속)

대표도 - 도2



노드와 자식 노드 사이의 에지를 포함하여 구성되고, 완전 이진 트리 구조를 가지는 뉴럴 트리 디코더(neural tree decoder)를 포함하며, 상기 뉴럴 트리 디코더는, 상기 계층적 변환기 인코더에서 추출한 상기 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측하는 것을 그 구성상의 특징으로 한다.

본 발명은 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기에 의한 분류 방법에 관한 것으로서, 보다 구체적으로는 컴퓨터로 구현되는 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기(ViT-NeT)에 의해 수행되는 분류 방법으로서, (1) 입력 이미지에서 주의 맵(attention map) 특징을 추출하는 계층적 변환기 인코더(hierarchical transformer encoder); 및 리프 노드를 포함하는 복수의 노드 및 상위 노드와 자식 노드 사이의 에지를 포함하여 구성되고, 완전 이진 트리 구조를 가지는 뉴럴 트리 디코더(neural tree decoder)를 포함하여 구성되는 ViT-NeT를 학습해 분류 모델을 생성하는 단계; 및 (2) 상기 분류 모델에 입력 이미지를 입력해 최종 클래스 레이블을 예측하는 단계를 포함하며, 상기 뉴럴 트리 디코더는, 상기 계층적 변환기 인코더에서 추출한 상기 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에 따르면, 비전 변환기(ViT) 인코더와 뉴럴 트리(NeT) 디코더를 결합해 ViT-NeT를 구성하되, 비전 변환기를 백본으로 사용해 주의 맵을 추출하고 이를 뉴럴 트리 디코더에 입력함으로써, 클래스 간 상관관계가 높고 클래스 내 상관관계가 낮은 세분화된 개체를 정확하게 분류할 수 있고, 트리 구조와 프로토타입을 통해 의사결정 과정을 설명하고 결과를 시각적으로 해석할 수 있어, 기존 앙상블 모델에서 발생하는 해석 가능성과 성능의 트레이드 오프를 해결할 수 있다.

(52) CPC특허분류

G06N 5/01 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345334317
과제번호	2021R1A6A3A13039438
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	박사과정생연구장려금지원
연구과제명	전방향 설명이 가능한 의사결정 신경망 구조 설계
기 여 율	1/1
과제수행기관명	계명대학교
연구기간	2021.06.01 ~ 2022.05.31

명세서

청구범위

청구항 1

뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기(ViT-NeT)로서,
 입력 이미지에서 주의 맵(attention map) 특징을 추출하는 계층적 변환기 인코더(hierarchical transformer encoder)(100); 및
 리프 노드를 포함하는 복수의 노드 및 상위 노드와 자식 노드 사이의 에지를 포함하여 구성되고, 완전 이진 트리 구조를 가지는 뉴럴 트리 디코더(neural tree decoder)(200)를 포함하며,
 상기 뉴럴 트리 디코더(200)는,
 상기 계층적 변환기 인코더(100)에서 추출한 상기 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 2

제1항에 있어서, 상기 계층적 변환기 인코더(100)는,
 시프트 윈도우(shifted window)를 사용해 윈도우 크기를 변경하여, 입력 이미지 내의 크기가 서로 다른 객체를 검출할 수 있는 주의 맵 특징을 추출하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 3

제2항에 있어서, 상기 계층적 변환기 인코더(100)는,
 스윈 변환기(swin transformer)를 백본 모델로 사용해 세밀한 분류와 이미지의 시각적 설명을 위한 상기 주의 맵 지도를 추출하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 4

제1항에 있어서, 상기 뉴럴 트리 디코더(200)는,
 프로토타입을 사용해 이미지 패치 중 차별적(discriminative) 특징을 찾아 경로설정 방향을 결정하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 5

제4항에 있어서, 상기 뉴럴 트리 디코더(200)는,
 각 노드에서 프로토타입과 이미지 패치 사이의 유사도(similarity)를 나타내는 라우팅 점수를 계산하고, 상기 라우팅 점수에 따라 경로설정 방향을 결정하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 6

제1항에 있어서, 상기 뉴럴 트리 디코더(200)의 에지는,

맥락 변환기 모듈을 이용해 각 위치의 이미지 패치에 전역 맥락(global context)을 집약하여 전역 특징이 강화된 이미지 패치를 자식 노드에 공급하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 7

제5항에 있어서, 상기 뉴럴 트리 디코더(200)의 리프 노드는,

클래스 확률을 예측하는 리프 예측 모듈에 의한 모든 리프 예측과 루트 노드에서 리프 노드까지의 경로별 에지 세트에 대한 누적 라우팅 점수를 이용해 최종 예측을 출력하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 8

제1항에 있어서, 상기 뉴럴 트리 디코더(200)는,

데이터셋의 복잡성에 따라 트리 깊이가 결정되는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기.

청구항 9

컴퓨터로 구현되는 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기(ViT-NeT)에 의해 수행되는 분류 방법으로서,

(1) 입력 이미지에서 주의 맵(attention map) 특징을 추출하는 계층적 변환기 인코더(hierarchical transformer encoder)(100); 및 리프 노드를 포함하는 복수의 노드 및 상위 노드와 자식 노드 사이의 에지를 포함하여 구성되고, 완전 이진 트리 구조를 가지는 뉴럴 트리 디코더(neural tree decoder)(200)를 포함하여 구성되는 ViT-NeT를 학습해 분류 모델을 생성하는 단계; 및

(2) 상기 분류 모델에 입력 이미지를 입력해 최종 클래스 레이블을 예측하는 단계를 포함하며,

상기 뉴럴 트리 디코더(200)는,

상기 계층적 변환기 인코더(100)에서 추출한 상기 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 10

제9항에 있어서, 상기 계층적 변환기 인코더(100)는,

시프트 윈도우(shifted window)를 사용해 윈도우 크기를 변경하여, 입력 이미지 내의 크기가 서로 다른 객체를 검출할 수 있는 주의 맵 특징을 추출하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 11

제10항에 있어서, 상기 계층적 변환기 인코더(100)는,

스윈 변환기(swin transformer)를 백본 모델로 사용해 세밀한 분류와 이미지의 시각적 설명을 위한 상기 주의 맵 지도를 추출하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 12

제9항에 있어서, 상기 뉴럴 트리 디코더(200)는,

프로토타입을 사용해 이미지 패치 중 차별적(discriminative) 특징을 찾아 경로설정 방향을 결정하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 13

제12항에 있어서, 상기 뉴럴 트리 디코더(200)는,

각 노드에서 프로토타입과 이미지 패치 사이의 유사도(similarity)를 나타내는 라우팅 점수를 계산하고, 상기 라우팅 점수에 따라 경로설정 방향을 결정하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 14

제9항에 있어서, 상기 뉴럴 트리 디코더(200)의 예지는,

맥락 변환기 모듈을 이용해 각 위치의 이미지 패치에 전역 맥락(global context)을 집약하여 전역 특징이 강화된 이미지 패치를 자식 노드에 공급하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 15

제13항에 있어서, 상기 뉴럴 트리 디코더(200)의 리프 노드는,

클래스 확률을 예측하는 리프 예측 모듈에 의한 모든 리프 예측과 루트 노드에서 리프 노드까지의 경로별 예지 세트에 대한 누적 라우팅 점수를 이용해 최종 예측을 출력하는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

청구항 16

제9항에 있어서, 상기 뉴럴 트리 디코더(200)는,

데이터셋의 복잡성에 따라 트리 깊이가 결정되는 것을 특징으로 하는, 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법.

발명의 설명**기술 분야**

[0001] 본 발명은 비전 변환기 및 그에 의한 분류 방법에 관한 것으로서, 보다 구체적으로는 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에 관한 것이다.

배경 기술

[0002] 최근 자연어 처리(natural language processing; NLP) 분야에서 성공적인 결과를 가져온 변환기(transformer) 구조는 컴퓨터 비전에도 적용되고 있다. NLP가 문장을 여러 단어로 나누고 변환기가 각 단어 간의 연관성을 학습할 수 있도록 하는 것과 같은 방식으로 비전 변환기(Vision Transformer; ViT)(Dosovitskiy et al., 2020)는 이미지 패치 간의 연관성을 학습한다. 결과적으로 ViT 구조가 CNN(Convolutional Neural Network) 구조보다 비전 분야에서 더 나은 성능을 보이는 것으로 입증되었다. 그러나 ViT 및 CNN 접근 방식은 블랙박스 학습 메커니즘을 기반으로 하므로 모델의 해석 가능성 측면에서 한계가 있다. 따라서 모델 해석 가능성, 결과 설명 및

모델 성능 정보에 대한 요구가 증가하고 있다.

[0004] 컴퓨터 비전 분야에서 해석 가능성은 크게 두 가지 접근 방식, 즉 예측 결과에 주의(attention)(예: 히트 맵 및 현저성(saliency))를 제공하는 접근 방식과 모델 자체의 해석 가능성(투명성)을 사용하는 접근 방식으로 나눌 수 있다. 첫째, 주의 기반 설명 모델(Samek et al., 2017; Selvaraju et al., 2017; Chattopadhyay et al., 2018; Itaya et al., 2021)은 예측 모델이 객체를 특정 클래스로 분류하는 이유를 설명하기 위해 이미지 픽셀에 현저성(saliency)을 제공한다. 그러나 이러한 방법들은 주의 맵에서 다시 모델에 대한 설명을 유추해야 하고 모델이 왜 이러한 픽셀을 강조하는지에 대한 의사결정 과정을 알 수 없다는 한계가 있다. 둘째, 해석 가능한 모델(Chen & Guestrin, 2016; Ke et al., 2017; Dorogush et al., 2018; Lundberg et al., 2020; Kim et al., 2021)은 모델 자체가 투명하거나 의사결정 단계를 이해할 수 있다. 모델의 해석 가능성은 복잡성 및 성능과 밀접한 관련이 있다. 복잡성의 관점에서 볼 때 원래 해석 가능한 구조로 설계된 의사결정 트리는 구성이 매우 단순하므로 DNN (Deep Neural Network)에 비해 해석 가능성 측면에서 이점이 있다. 세분화된 이미지에서 높은 성능을 유지하면서 DNN의 낮은 해석 가능성을 극복하기 위해 일부 알고리즘(Chen et al., 2019a; Ji et al., 2020; Wan et al., 2021; Nauta et al., 2021)은 DNN을 의사결정 트리와 결합하려고 시도했다. 이러한 알고리즘은 분류 과정에 대한 설명뿐만 아니라 분류 이유에 대해 사용자에게 주의 맵(attention map)을 제공한다. 그러나 이러한 해석 가능한 모델은 의사결정 프로세스를 설명하는 데 여전히 한계가 있으며 해석 가능성과 성능 사이에는 트레이드 오프가 있다. 예를 들어, ProtoTree(Nauta et al., 2021)는 DNN과 의사결정 트리의 앙상블을 사용하여 가장 세밀한 이미지 분류 성능을 달성했으나, 앙상블 모델의 학습 복잡성으로 인해 결국 해석 가능성이 다시 낮아졌다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 비전 변환기(ViT) 인코더와 뉴럴 트리(NeT) 디코더를 결합해 ViT-NeT를 구성하되, 비전 변환기를 백본으로 사용해 주의 맵을 추출하고 이를 뉴럴 트리 디코더에 입력함으로써, 클래스 간 상관관계가 높고 클래스 내 상관관계가 낮은 세분화된 개체를 정확하게 분류할 수 있고, 트리 구조와 프로토타입을 통해 의사결정 과정을 설명하고 결과를 시각적으로 해석할 수 있어, 기존 앙상블 모델에서 발생하는 해석 가능성과 성능의 트레이드 오프를 해결할 수 있는, 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법을 제공하는 것을 그 목적으로 한다.

과제의 해결 수단

[0006] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기는,
 [0007] 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기(ViT-NeT)로서,
 [0008] 입력 이미지에서 주의 맵(attention map) 특징을 추출하는 계층적 변환기 인코더(hierarchical transformer encoder); 및
 [0009] 리프 노드를 포함하는 복수의 노드 및 상위 노드와 자식 노드 사이의 에지를 포함하여 구성되고, 완전 이진 트리 구조를 가지는 뉴럴 트리 디코더(neural tree decoder)를 포함하며,
 [0010] 상기 뉴럴 트리 디코더는,
 [0011] 상기 계층적 변환기 인코더에서 추출한 상기 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측하는 것을 그 구성상의 특징으로 한다.

[0013] 바람직하게는, 상기 계층적 변환기 인코더는,
 [0014] 시프트 윈도우(shifted window)를 사용해 윈도우 크기를 변경하여, 입력 이미지 내의 크기가 서로 다른 객체를 검출할 수 있는 주의 맵 특징을 추출할 수 있다.

- [0016] 더욱 바람직하게는, 상기 계층적 변환기 인코더는,
- [0017] 스윈 변환기(swin transformer)를 백본 모델로 사용해 세밀한 분류와 이미지의 시각적 설명을 위한 상기 주의 맵 지도를 추출할 수 있다.
- [0019] 바람직하게는, 상기 뉴럴 트리 디코더는,
- [0020] 프로토타입을 사용해 이미지 패치 중 차별적(discriminative) 특징을 찾아 경로설정 방향을 결정할 수 있다.
- [0022] 더욱 바람직하게는, 상기 뉴럴 트리 디코더는,
- [0023] 각 노드에서 프로토타입과 이미지 패치 사이의 유사도(similarity)를 나타내는 라우팅 점수를 계산하고, 상기 라우팅 점수에 따라 경로설정 방향을 결정할 수 있다.
- [0025] 바람직하게는, 상기 뉴럴 트리 디코더의 예지는,
- [0026] 맥락 변환기 모듈을 이용해 각 위치의 이미지 패치에 전역 맥락(global context)을 집약하여 전역 특징이 강화된 이미지 패치를 자식 노드에 공급할 수 있다.
- [0028] 더더욱 바람직하게는, 상기 뉴럴 트리 디코더의 리프 노드는,
- [0029] 클래스 확률을 예측하는 리프 예측 모듈에 의한 모든 리프 예측과 루트 노드에서 리프 노드까지의 경로별 예지 세트에 대한 누적 라우팅 점수를 이용해 최종 예측을 출력할 수 있다.
- [0031] 바람직하게는, 상기 뉴럴 트리 디코더는,
- [0032] 데이터셋의 복잡성에 따라 트리 깊이가 결정될 수 있다.
- [0034] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기에 의한 분류 방법은,
- [0035] 컴퓨터로 구현되는 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기(ViT-NeT)에 의해 수행되는 분류 방법으로서,
- [0036] (1) 입력 이미지에서 주의 맵(attention map) 특징을 추출하는 계층적 변환기 인코더(hierarchical transformer encoder); 및 리프 노드를 포함하는 복수의 노드 및 상위 노드와 자식 노드 사이의 예지를 포함하여 구성되고, 완전 이진 트리 구조를 가지는 뉴럴 트리 디코더(neural tree decoder)를 포함하여 구성되는 ViT-NeT를 학습해 분류 모델을 생성하는 단계; 및
- [0037] (2) 상기 분류 모델에 입력 이미지를 입력해 최종 클래스 레이블을 예측하는 단계를 포함하며,
- [0038] 상기 뉴럴 트리 디코더는,
- [0039] 상기 계층적 변환기 인코더에서 추출한 상기 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측하는 것을 그 구성상의 특징으로 한다.
- [0041] 바람직하게는, 상기 계층적 변환기 인코더는,
- [0042] 시프트 윈도우(shifted window)를 사용해 윈도우 크기를 변경하여, 입력 이미지 내의 크기가 서로 다른 객체를 검출할 수 있는 주의 맵 특징을 추출할 수 있다.

- [0044] 더욱 바람직하게는, 상기 계층적 변환기 인코더는,
- [0045] 스윈 변환기(swin transformer)를 백본 모델로 사용해 세밀한 분류와 이미지의 시각적 설명을 위한 상기 주의 맵 지도를 추출할 수 있다.
- [0047] 바람직하게는, 상기 뉴럴 트리 디코더는,
- [0048] 프로토타입을 사용해 이미지 패치 중 차별적(discriminative) 특징을 찾아 경로설정 방향을 결정할 수 있다.
- [0050] 더욱 바람직하게는, 상기 뉴럴 트리 디코더는,
- [0051] 각 노드에서 프로토타입과 이미지 패치 사이의 유사도(similarity)를 나타내는 라우팅 점수를 계산하고, 상기 라우팅 점수에 따라 경로설정 방향을 결정할 수 있다.
- [0053] 바람직하게는, 상기 뉴럴 트리 디코더의 예지는,
- [0054] 맥락 변환기 모듈을 이용해 각 위치의 이미지 패치에 전역 맥락(global context)을 집약하여 전역 특징이 강화된 이미지 패치를 자식 노드에 공급할 수 있다.
- [0056] 더더욱 바람직하게는, 상기 뉴럴 트리 디코더의 리프 노드는,
- [0057] 클래스 확률을 예측하는 리프 예측 모듈에 의한 모든 리프 예측과 루트 노드에서 리프 노드까지의 경로별 예지 세트에 대한 누적 라우팅 점수를 이용해 최종 예측을 출력할 수 있다.
- [0059] 바람직하게는, 상기 뉴럴 트리 디코더는,
- [0060] 데이터셋의 복잡성에 따라 트리 깊이가 결정될 수 있다.

발명의 효과

- [0061] 본 발명에서 제안하고 있는 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에 따르면, 비전 변환기(ViT) 인코더와 뉴럴 트리(NeT) 디코더를 결합해 ViT-NeT를 구성하되, 비전 변환기를 백본으로 사용해 주의 맵을 추출하고 이를 뉴럴 트리 디코더에 입력함으로써, 클래스 간 상관관계가 높고 클래스 내 상관관계가 낮은 세분화된 개체를 정확하게 분류할 수 있고, 트리 구조와 프로토타입을 통해 의사결정 과정을 설명하고 결과를 시각적으로 해석할 수 있어, 기존 앙상블 모델에서 발생하는 해석 가능성과 성능의 트레이드 오프를 해결할 수 있다.

도면의 간단한 설명

- [0062] 도 1은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기의 개념을 설명하기 위해 도시한 도면.
- 도 2는 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기의 구성을 도시한 도면.
- 도 3은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기의 학습 알고리즘을 도시한 도면.
- 도 4는 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기에 의한 분류 방법의 흐름을 도시한 도면.
- 도 5는 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법과 최신 모델의 CUB-200-2011 데이터셋에 대한 정확도 비교 결과를 나타낸 도면.
- 도 6은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방

법과 최신 모델의 Stanford Dogs 데이터셋에 대한 정확도 비교 결과를 나타낸 도면.

도 7은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법과 최신 모델의 Stanford Cars 데이터셋에 대한 정확도 비교 결과를 나타낸 도면.

도 8은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 트리 깊이에 따른 정확도를 나타낸 도면.

도 9는 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 트리 깊이가 5일 때 특정 의사결정 경로에서의 프로토타입 반응을 나타낸 도면.

도 10은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법의 풀링 방법에 따른 실험 결과를 나타낸 도면.

도 11은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 맥락 변환기 모듈에 따른 실험 결과를 나타낸 도면.

도 12는 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 무작위로 샘플링된 이미지에 대한 순차적 의사결정을 시각화해 도시한 도면.

발명을 실시하기 위한 구체적인 내용

[0063] 이하, 첨부된 도면을 참조하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명을 용이하게 실시할 수 있도록 바람직한 실시예를 상세히 설명한다. 다만, 본 발명의 바람직한 실시예를 상세하게 설명함에 있어, 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다. 또한, 유사한 기능 및 작용을 하는 부분에 대해서는 도면 전체에 걸쳐 동일한 부호를 사용한다.

[0065] 덧붙여, 명세서 전체에서, 어떤 부분이 다른 부분과 ‘연결’되어 있다고 할 때, 이는 ‘직접적으로 연결’되어 있는 경우뿐만 아니라, 그 중간에 다른 소자를 사이에 두고 ‘간접적으로 연결’되어 있는 경우도 포함한다. 또한, 어떤 구성요소를 ‘포함’한다는 것은, 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있다는 것을 의미한다.

[0067] 본 발명은 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기(Vision Transformer Neural Tree; ViT-NeT) 및 그에 의한 분류 방법에 관한 것으로서, 본 발명의 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법은, 메모리 및 프로세서를 포함한 하드웨어에서 기록되는 소프트웨어로 구성될 수 있다. 예를 들어, 본 발명의 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법은, 개인용 컴퓨터, 노트북 컴퓨터, 서버 컴퓨터, PDA, 스마트폰, 태블릿 PC 등에 저장 및 구현될 수 있다.

[0069] 본 발명은 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기에 관한 것으로, ViT-NeT는 우수한 FGVC(fine-grained visual categorization)를 지원하는 동시에 시각적 설명과 모델 해석 가능성을 모두 제공할 수 있다. ViT-NeT는 ViT 인코더와 NeT 디코더(Soft Neural Tree Decoder)의 조합으로 구성될 수 있다. 여기서, ViT의 역할은 고품질 로컬 및 글로벌 특징과 주의(attention) 정보를 사용하여 높은 수준의 분류 성능을 달성하는 것이고, 뉴럴 트리의 역할은 ViT의 의사결정 프로세스를 해석하고 이미지를 계층적으로 라우팅하는 판별 디코더 역할을 하는 것이다. 따라서 ViT-NeT는 기존 양상블 모델에서 발생하는 해석 가능성과 성능의 트레이드 오프를 해결할 수 있다. 즉, ViT-Net은 뉴럴 트리 디코더를 사용하여 ViT에 고유한 해석 가능성을 부여하는 새로운 방법으로서, 주어진 영상에 대한 순차적인 의사결정 절차의 해석을 지원할 수 있다.

[0071] 도 1은 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기의 개념을 설명하기 위해 도시한 도면이다. 도 1에 도시된 ViT-NeT의 대략적인 동작 구조를 살펴보면, 뉴럴 트리 디코더의 노드별로 어떤 이미지 패치가 주목을 받는지, 이미지 패치가 노드별로 어떻게 분기되는지, 마지막으로 세밀한 분류 결과를 알 수 있다. 보다 구체적으로, 각 트리 노드에서 영덩이, 녹색 목구멍, 꼬리 덮깃(tail covert)에 각각 주

의가 집중되며, 이미지는 특정 노드 경로를 통해 “새”로 분류될 수 있다. ViT-NeT는 미세한 시각적 분류를 위해 모든 ViT 방법 및 데이터셋과 호환될 수 있다.

- [0073] 본 발명은 ViT와 뉴럴 의사결정 트리(neural decision tree)를 결합하여 미세한 이미지에 대한 높은 해석 가능성 및 분류 성능을 달성한 최초의 연구이다. 본 발명에서 제안한 ViT-Net의 기여도는 다음과 같다.
- [0075] - 기존 해석 가능한 모델과 달리 SOTA(state-of-the-art) CNN 및 ViT와 경쟁적으로 세분화된 이미지 분류에 대한 높은 성능을 달성하고 성능과 해석 가능성 간의 절충 없이 분석할 수 있다.
- [0077] - 바닐라 ViT의 주의 가중치는 주의 맵(Attention Map)과 같은 이미지 해석 가능성만 지원하는 반면, 본 발명은 의사결정 모델 자체를 투명하게 만들어 의사결정 과정과 이미지 설명을 동시에 가능하게 한다.
- [0079] - 분자 프로토타이핑을 통해 뉴럴 트리 노드에서 데이터의 긍정적인 요소와 부정적인 요소를 동시에 해석할 수 있다.
- [0081] 이하에서는, 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기를 구성하는 비전 변환기(ViT)와 뉴럴 트리(NeT)의 기존 연구에 대해 각각 설명하도록 한다.
- [0083] 먼저, ViT를 사용한 FGVC 연구를 살펴보도록 한다. 최근까지 시각적 분류는 딥러닝 연구의 가장 기본적인 분야였지만, SOTA 모델의 성능은 이미 한계에 도달했다. 또한 시각적으로 유사한 동식물을 구별하는 것과 같은 세밀한 분류는 인간과 기계 성능의 한계를 뛰어넘는 어려운 문제이다. 이 제한은 대부분의 이미지가 클래스 간 상관관계가 높고 클래스 내 상관관계가 낮아서 발생한다. 최근 다양한 ViT 기반 방법(Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021)이 CNN을 능가하고 SOTA 성능을 입증함에 따라 ViT가 FGVC에 적용되기 시작했다(He et al., 2021; Zhang et al. 2021). 예를 들어, TransFG(He 등, 2021년)는 주의 기반 패치 증류와 대조적 손실(contrastive loss)을 이용하여 강력한 국소 이미지 패치와 차별적 영역을 추출했다. AFTrans(Zhang et al., 2021)는 가중치 매개변수를 공유하는 삼 아키텍처를 사용하여 작동하는 선택적 주의 수집 모듈을 제안했다. 단, 변환기의 주의 모듈은 국소적인 고차별 부분에서 추출된다는 암묵적인 가정에 기초하여 적용되기 때문에 학습 과정 중에 생성되는 주의 맵의 신뢰성이 모호하다는 한계가 있다.
- [0085] 다음으로, ViT의 시각적 설명에 관한 연구를 살펴보도록 한다. CNN의 성능이 향상됨에 따라, 실제 적용에 필요한 신뢰성을 확보하기 위해 시각적인 설명에 대한 연구가 활발히 진행되고 있다. 지금까지 컴퓨터 비전에 대한 시각적 설명 기여는 주로 CNN에 초점을 맞췄다(Binder et al., 2016; Ribeiro et al., 2016; Selvaraju et al., 2017; Shrikumar et al., 2017; Lundberg & Lee, 2017; Chattopadhyay et al., 2018; Itaya et al., 2021). ViT가 새로운 학습 패러다임으로 떠오르고 있지만, 시각적 설명에 대한 실질적인 연구는 거의 이루어지지 않았다. ViT의 출력을 설명하는 일반적인 방법은 주의 맵을 시각화하는 것이다. 그러나 자기 주의(self-attention)를 통해 작성된 주의 맵은 쿼리와 키만을 사용하여 적용되기 때문에 모델의 전체적인 운용이 고려되지 않는다는 단점이 있다. 따라서 진정한 시각적 설명을 위해서는 이미지 해석 가능성과 동시에 의사결정 과정을 설명할 수 있는 새로운 모델이 필요하다.
- [0087] 한편, 뉴럴 트리(NeT)를 사용한 모델 해석성을 살펴보도록 한다. 의사결정 트리(Decision Tree)는 특정 태스크를 단일 특징 수준에서 해결하는 강력한 방법으로, 의사결정 트리 구조에는 본질적인 해석성이 있지만 컴퓨터 비전 태스크에 적합하지 않다. 이를 극복하기 위해 많은 연구(Hinton 등, 2015; Kotschierer 등, 2015; Balestrieri, 2017; Tanno 등, 2019; Ji 등, 2020; Alaniz 등, 2021; Wan 등; Nauta 등, 2021)가 다양한 시각적 과제와 차별화된 신경 호환성을 도입했다. ANTs(Tanno et al., 2019)는 학습 가능한 트리 구성요소, 즉 노

드(nodes), 에지(edges) 및 리프(leaves)를 사용하여 적응적으로 구성 가능한 뉴럴 트리 메커니즘을 이용한다. 그러나 해석은 제한적이며 대규모 데이터셋에는 적합하지 않다. ProtoTree (Nauta et al., 2021)에서는 프로토타입으로 의사결정을 시각화할 수 있는 해석 가능한 트리를 구축하기 위한 새로운 방법이 도입되었다. 그럼에도 불구하고, 최상의 성능은 그것이 형성된 앙상블일 때만 발생하며, 트리의 앙상블은 결국 해석성을 감소시키고 학습 계획이 복잡하다는 점에서 단점이 있다.

[0089] 진술한 바와 같은 문제를 해결하기 위해, 본 발명에서는 모델의 해석 가능성과 정확성 사이의 더 나은 균형을 달성하기 위해 계층적 변환기 인코더 (hierarchical transformer encoder)를 뉴럴 트리 디코더 (neural tree decoder)와 결합하여, 새롭고 구조적으로 해석 가능하며 시각적으로 설명 가능한 ViT-NeT 모델을 제안한다.

[0091] 도 2는 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기를 설명하기 위해 도시한 도면이다. 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기는, 계층적 변환기 인코더(100) 및 뉴럴 트리 디코더(200)를 포함하며, 뉴럴 트리 디코더(200)는 계층적 변환기 인코더(100)에서 추출한 주의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측할 수 있다.

[0093] 즉, 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기는, ViT 방법 중 계층적 비전 변환기를 특징 표현을 위한 백본으로 사용할 수 있다. 여기서, 계층적 비전 변환기는 시프트 창을 사용하여 이미지 내의 작은 개체와 큰 개체의 변화를 동시에 고려할 수 있다. 다음으로, 국지적 수준의 특징 프로토타이핑(Chen et al., 2018; Nauta et al., 2021)을 보다 인간 친화적인 접근법으로써, 의사결정 과정을 쉽게 해석하고 시각적 설명을 가능하게 하는 동시에 세밀한 분류 성능을 향상시키도록 설계되었다.

[0095] 이하에서는, 도 2를 참조하여 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기를 구성하는 각 구성요소에 대해 상세히 설명하도록 한다.

[0097] 계층적 변환기 인코더(100)는, 입력 이미지에서 주의 맵(attention map) 특징을 추출할 수 있다. 보다 구체적으로, 계층적 변환기 인코더(100)는, 시프트 윈도우(shifted window)를 사용해 윈도우 크기를 변경하여, 입력 이미지 내의 크기가 서로 다른 객체를 검출할 수 있는 주의 맵 특징을 추출할 수 있다. 또한, 계층적 변환기 인코더(100)는, 스윈 변환기(swin transformer)를 백본 모델로 사용해 세밀한 분류와 이미지의 시각적 설명을 위한 주의 맵 지도를 추출할 수 있다.

[0099] 즉, 여러 ViT 중 스윈 변환기(Swin Transformer)(Liu 등, 2021년)는 시프트 윈도우 메커니즘을 사용하여 유도 바이어스 및 계산 복잡성 등 초기 ViT의 다양한 문제를 해결하기 위한 강력한 표현을 유도한다. 기존 변환기에 사용되는 단일 창 접근법은 CNN과 달리 유도 바이어스를 유지해야 하는 다양한 교란된 영상에서 몇 가지 약점을 보인다. 전통적인 CNN이 많은 경우에서 강건성을 보장하는 것과 같은 방식으로, 스윈 변환기는 CNN 프레임워크에서 특징 피라미드와 유사한 계층적 특징 인코딩을 상속한다. 스윈 변환기에서는 시프트 윈도우를 사용하여 윈도우 크기를 변경하고 여러 패치를 셀프 어텐션(self-attention)으로 처리한다. 윈도우에 포함된 패치의 크기가 서로 다르기 때문에 작은 물체와 큰 물체를 쉽게 검출할 수 있는 주의 맵이 작성될 수 있다. ViT 중 하나인 바닐라 변환기의 셀프 어텐션은 입력 이미지에 2차 계산 횟수를 적용하는 반면, 스윈 변환기는 이미지에 선형 계산을 수행하므로, 상대적으로 적은 수의 계산으로 모델의 크기를 증가시킬 수 있을 뿐만 아니라 추론 속도를 높일 수 있는 장점이 있다. 계층적 변환기 인코더(100)는, 스윈 변환기를 백본 모델로 사용하여 세밀한 분류와 이미지의 시각적 설명을 위한 주의 맵 특징을 추출할 수 있다.

[0101] 뉴럴 트리 디코더(200)는, 리프 노드를 포함하는 복수의 노드 및 상위 노드와 자식 노드 사이의 에지를 포함하여 구성되고, 완전 이진 트리(perfect binary tree) 구조를 가지며, 계층적 변환기 인코더(100)에서 추출한 주

의 맵 특징을 입력으로 하여 최종 클래스 레이블을 예측할 수 있다. 즉, 본 발명에서는, 정확성과 해석 가능성 사이의 더 나은 균형을 달성하기 위해, 완전 이진 트리와 차별화된 경로설정 메커니즘(routing mechanism)을 사용하는 뉴럴 트리 디코더(200)를 제안한다.

[0103] 보다 구체적으로, 도 2의 (b)에 도시된 바와 같이, 뉴럴 트리 디코더(200)는 노드 $N(\cdot)$ (흰색 원으로 표시), 리프 노드 $L(\cdot)$ (회색 원으로 표시) 및 상위 노드 i 와 자식 노드(하위 노드) j 사이의 에지 $\varepsilon_{i,j}(\cdot)$ (연두색 원으로 표시) 세트로 구성될 수 있다. 뉴럴 트리 디코더(200)는 완전 이진 트리를 사용하기 때문에 각 내부 노드에는 $N_{2 \times i}$ 및 $N_{2 \times i+1}$ 두 개의 자식 노드가 있다. 계층적 변환기 인코더(100)의 출력 $\mathbf{z}_1 = f(\mathbf{x}; \Theta)$ 가 주어졌을 때, 뉴럴 트리 디코더(200)는 소프트 리셋 방식으로 최종 클래스 레이블 "y"를 예측한다.

[0105] 도 3은 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기의 학습 알고리즘을 도시한 도면이다. 도 3에 도시된 바와 같이, 학습 알고리즘은 ViT-NeT를 통과하는 재귀 트리를 통해 학습 절차를 도입할 수 있다.

[0107] ViT-NeT 모델의 제안된 NeT는 특히 클래스 간 상관관계가 유사하고 클래스 내 상관관계가 다른 개체를 분류하는 것을 목적으로 한다. 글로벌 분류는 비교적 단순한 글로벌 신호를 기반으로 적용할 수 있지만 세밀한 분류는 고도로 국지화된 세분화된 범주의 구별 영역을 필요로 한다. 따라서 본 발명에서는 작은 식별 특징을 포착하기 위해 축소 이미지 패치에 프로토타입을 적용할 수 있다.

[0109] 뉴럴 트리 디코더(200)는, 프로토타입을 사용해 이미지 패치 중 차별적(discriminative) 특징을 찾아 경로설정 방향을 결정할 수 있다. 즉, NeT는 프로토타입을 사용하여 이미지 패치 중 차별적 영역을 찾는다. 차별적 특징이 샘플링되면 경로설정 방향이 결정되어 구별할 수 있는 라우팅 모듈(differentiable routing module)의 자식 노드에 샘플이 전송된다.

[0111] 보다 구체적으로, 뉴럴 트리 디코더(200)는, 각 노드에서 프로토타입과 이미지 패치 사이의 유사도(similarity)를 나타내는 라우팅 점수를 계산하고, 라우팅 점수에 따라 경로설정 방향을 결정할 수 있다. 각 내부 노드는 학습 가능한 프로토타입(trainable prototype) $\mathcal{P}_i \in \mathbb{R}^{H_i \times W_i \times D}$ 에 대응한다. 각 프로토타입은 다음 수학적 1에 의해 계산되는 이미지 패치 $\mathbf{z} \in \mathbb{R}^{H \times W \times D}$ 에 대한 라우팅 점수(routing score) $N_i \in [0, 1]$ 의 제곱 L^2 을 측정하는 데 사용되는 학습 가능한 텐서이다.

수학적 1

[0113]
$$\mathcal{N}(\mathbf{z}_i) = \max_{\tilde{\mathbf{z}} \in \text{patches}(\mathbf{z}_i)} \log((\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + 1) / (\|\tilde{\mathbf{z}} - \mathcal{P}_i\|_2^2 + \epsilon))$$

[0115] i 번째 내부 노드의 라우팅 점수 $N(\mathbf{z}_i)$ 은 가장 가까운 응축 이미지 패치(condensed image patch) $\tilde{\mathbf{z}} \in \mathbb{R}^{1 \times D}$ 와 프로토타입 $\mathcal{P}_i$ 와의 유사도(similarity)를 나타낸다. 각 노드에서 프로토타입과 이미지 패치 간의 유사도를 정의하기 위해 각 프로토타입이 $\mathbf{z}$ 위의 컨볼루션 커널로 작동하며 프로토타입과 현재 $\mathbf{z}$ 사이의 거리를 도출하는 컨볼루션 연산을 사용한다. 거리에서의 유사도를 유도하기 위해 수학적 1에 있는 로그 유사도 척도를 사용한다. 수학적 1은 거리가 수용 패치(receptive patch)에 가까울 때 유사도가 높다. 수학적 1에 따라 각 자식 노드에 대한 라우팅 점수를 각각 $\mathcal{R}_{i,2 \times i}(\mathbf{z}_i) = [\mathcal{N}(\mathbf{z}_i)]_0^1$ 및 $\mathcal{R}_{i,2 \times i+1}(\mathbf{z}_i) = 1 - [\mathcal{N}(\mathbf{z}_i)]_0^1$ 로 정의한다.

[0117] 한편, 뉴럴 트리 디코더(200)의 예지는, 맥락 변환기 모듈(contextual transformer module; CTM)을 이용해 각 위치의 이미지 패치에 전역 맥락(global context)을 집약하여 전역 특징이 강화된 이미지 패치를 자식 노드에 공급할 수 있다. 여기서, 맥락 변환기 모듈은 ViT-NeT 모델이 차별적 패치를 캡처하도록 강제하기 위해 사용될 수 있다. 경험적 수용장(receptive field)이 이론적 수용장보다 훨씬 작다는 사실에 따르면, 차별적 표현은 더 큰 수용장으로 형성되어야 한다. 이와 관련하여, 모든 이산화된 이미지 패치에서 트리 예지로 전역 특징(global features)을 강화하는 데 초점을 맞추었다.

[0119] 도 2의 (c)에 도시된 바와 같이, 맥락 변환기 모듈(CTM)은 전역 맥락을 각 위치의 패치에 집약함으로써 객체에 대한 보다 나은 설명을 제공할 수 있다. 보다 구체적으로, 주어진 이미지 패치는 사이드 채널 차원에 따라 1차원까지 압축되고, 패치 차원에 소프트맥스가 적용될 수 있다. 이 축약된 맥락 매트릭스에 주어진 이미지 패치를 곱하여 전역 맥락 정보를 할당한다. 이어서 매트릭스가 투영된 다음 GELU와 레이어 정규화(LN)가 포함된 두 개의 완전 연결 레이어(fully connected layer; FC)가 나타난다. 마지막으로 다음 수학적 2를 이용해 도출되는 강화된 맥락 이미지 패치 $\mathbf{z}_j$ 를 대응하는 자식/리프 노드에 공급한다.

수학적 2

[0121]
$$\mathbf{z}_j = \mathcal{E}_{i,j}(\mathbf{z}_i) = \text{CTM}(\mathbf{z}_i)$$

[0123] 한편, 뉴럴 트리 디코더(200)의 리프 노드는, 클래스 확률을 예측하는 리프 예측 모듈에 의한 모든 리프 예측과 루트 노드에서 리프 노드까지의 경로별 예지 세트에 대한 누적 라우팅 점수를 이용해 최종 예측을 출력할 수 있다. 보다 구체적으로, 뉴럴 트리 디코더(200) 내의 각 리프는 학습할 필요가 있는 K 클래스에 대한 클래스 확률을 예측하기 위한 리프 예측 모듈 $L(\cdot)$ 에 대응할 수 있다. $p(\mathbf{z}_1)$ 는 루트 노드에서 1번째 리프 노드로 통과하는 특정 경로 p_1 내의 예지 세트에서, 계층적 변환기 인코더(100)의 출력 $\mathbf{z}_1$ 의 누적 라우팅 점수라고 하자. 누적 라우팅 점수는, 다음 수학적 3에 의해 계산될 수 있다.

수학적 3

[0125]
$$\rho(\mathbf{z}_i) = \prod_{(i,j) \in p_i} \mathcal{R}_{i,j}(\mathcal{E}_{i,j}(\mathbf{z}_i))$$

[0127] 각 리프 예측 모듈은 도 2의 (d)와 같이 글로벌 평균 풀링(GAP), 글로벌 최대 풀링(GMP), 공유 가중치를 가진 글로벌 완전 연결 레이어(GFC), 로컬 완전 연결 레이어(LFC) 및 레이어 정규화(LN)를 통해 구성될 수 있다. 리프 예측 모듈의 공식은 다음 수학적 4와 같다.

수학적 4

[0129]
$$\mathcal{L}(\mathbf{z}_l, x) = \text{LN}(\text{LFC}(\text{GMP}(\mathbf{z}_l)) + \text{GFC}(\text{GAP}(f(x; \theta))))$$

[0131] 수학적식 4에서 리프 예측 모듈은 주어진 강화된 맥락 이미지 패치와 LFC를 사용하여 로컬 측면(local view)을 학습하고, GFC를 사용하여 글로벌 맥락 측면을 학습할 수 있다.

[0133] 최종 예측 $\hat{y}$ 은 다음 수학적식 5와 같이 모든 리프 예측 L에 누적 라우팅 점수 ρ_l 을 곱한 합계로 계산될 수 있다.

수학적식 5

$$\hat{y} = \sum_{l \in L} \sigma(\mathcal{L}(\mathbf{z}_l, x)) \cdot \rho_l(\mathbf{z}_l)$$

[0137] 최종 예측은 실측(ground truth) 레이블 y 를 이용해 음의 로그 우도 손실(negative logarithmic likelihood loss)을 사용해 최적화될 수 있다.

[0139] 도 4는 본 발명의 일실시예에 따른 뉴럴 트리 디코더를 이용한 해석 가능한 비전 변환기(200)에 의한 분류 방법의 흐름을 도시한 도면이다. 도 4에 도시된 바와 같이, 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기에 의한 분류 방법은, 계층적 변환기 인코더(100) 및 뉴럴 트리 디코더(200)를 포함하여 구성되는 ViT-NeT를 학습해 분류 모델을 생성하는 단계(S100) 및 분류 모델에 입력 이미지를 입력해 최종 클래스 레이블을 예측하는 단계(S200)를 포함하여 구현될 수 있다.

[0141] 각각의 단계들과 관련된 상세한 내용들은, 앞서 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기와 관련하여 충분히 설명되었으므로, 상세한 설명은 생략하기로 한다.

[0143] 실험 셋업

[0144] 전술한 바와 같은 ViT-NeT의 성능 확인을 위해, 정량 분석을 수행한 후 애플레이션 연구(Ablation Study)와 정성 분석을 수행했다.

[0146] 데이터셋으로는, CUB-200-2011(Wah 등, 2011), Stanford Cars(Krause 등, 2013), Stanford Dogs(Khosla 등, 2011)의 3개의 FGVC 데이터셋에 대해 ViT-NeT를 평가하고 정확성과 해석성 측면에서 이전 SOTA 모델과 비교했다.

[0148] ViT-NeT는 PyTorch에서 구현되었다. 계층형 변환기 인코더 f 는 ImageNet-22K를 사용하여 사전 학습된 Swin 변환기 레이어(Liu 등, 2021년)를 사용했다. 먼저 입력 이미지의 크기를 학습용에 대해 랜덤 크로핑(random cropping), 테스트용에 대해 센터 크로핑(center reopping)을 각각 적용하여, 448×448 픽셀 해상도로 조정했다. 이미지를 4개의 작은 패치로 분할하고 그룹 윈도우 크기를 14로 설정했다. AdamW 옵티마이저는 0.9의 모멘텀으로 적용되었다. 학습률은 CUB-200-2011의 경우 $2e-5$, Stanford Dogs의 경우 $2e-4$, Stanford Cars의 경우 $2e-3$ 로 초기화되었다. 배치 사이즈는 16으로 설정했다. 학습과 테스트는 4대의 NVIDIA Tesla V100 32GB GPU와 APEX를 사용하여 실시했다.

[0150] 정량 분석 1 - CUB-200-2011 데이터셋에 대한 평가

[0151] CUB-200-2011(Wah 등, 2011) 데이터셋은 학습 이미지 5,994개와 테스트 이미지 5,794개로 200개 클래스에

11,788개의 이미지로 구성되어 있다. 도 5는 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법과 최신 모델의 CUB-200-2011 데이터셋에 대한 정확도(accuracy) 비교 결과를 나타낸 도면이다. 도 5에 도시된 바와 같이, 본 발명의 ViT-NeT는 비교 대상 중 가장 우수한 AFTrans보다 나은 결과를 보여주며, CUB-200-2011 데이터셋에 대한 모든 CNN 기반 방법보다 정확도가 높다.

[0153] ViT-NeT의 백본 네트워크를 DeiT-B (base data-efficient image transformers)(Touvron 등, 2021년)로 대체하면 SwinT-B (base Swin transformer)에 비해 성능이 1.5% 저하된다. CUB-200-2011 데이터셋의 경우 제안된 NeT를 SwinT와 결합할 때 기본 변환기뿐만 아니라 CNN 기반 방법과 비교하여 가장 세밀한 분류 성능을 보인다는 것을 확인할 수 있다.

[0155] 정량 분석 2 - Stanford Dogs 데이터셋에 대한 평가

[0156] Stanford Dogs 데이터셋(Khosla et al., 2011)는 학습용 12,000개의 이미지와 테스트용 8,580개의 이미지로 구성된 20,580개의 다른 품종의 100개의 세분화된 데이터셋이다. 도 6은 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법과 최신 모델의 Stanford Dogs 데이터셋에 대한 정확도 비교 결과를 나타낸 도면이다. 도 6에 도시된 바와 같이, 본 발명의 ViT-NeT는 대부분의 비교 방법, 특히 CNN 백본보다 더 나은 성능을 보였다. 또한, ViT-NeT 모델에서 SwinT-B를 백본으로 사용하는 경우 DeiT-B 백본에 비해 상위 1의 정확도가 3.3% 정도 떨어진다. 이는 배경화면이 복잡한 이미지에서 SwinT-B의 분류 성능이 다소 떨어진다는 문제 때문이다. 다만, DeiT-B와 조합하면 DeiT-B와 SOTA AFTrans를 크게 웃돈다. 그 결과, 본 발명의 NeT는 ViT의 유형과 관계없이 뉴럴 트리 및 라우팅 메커니즘과 함께 전달된 응축 이미지 패치를 사용함으로써 세밀한 분류 성능을 향상시킬 수 있음을 입증했다.

[0158] 정량 분석 3 - 스탠포드 자동차 데이터셋에 대한 평가

[0159] Stanford Cars 데이터셋(Krause et al., 2013)에는 학습용 8,144개의 이미지와 테스트용 8,041개의 이미지를 포함하여 196개 클래스에 16,185개의 이미지가 포함되어 있다. 도 7은 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법과 최신 모델의 Stanford Cars 데이터셋에 대한 정확도 비교 결과를 나타낸 도면이다. 도 7에 도시된 바와 같이, SwinT-B를 기반으로 제안된 ViTNeT는 95.0%의 최고 정확도를 보였다. 이 정확도는 ViT 기반 방법 중 SOTA 성능을 달성한 AFTrans (Zhang et al., 2021)의 정확도와 동일하다. CNN 기반 SOTA 방법 중 API-Net (Zhuang et al., 2020)은 차별적 특징을 이용하기 위해 생성적 예측 전략과 함께 복잡한 DenseNet-161 및 반-섭동(anti-perturbation) 추론 설계를 채택하여 95.3%의 최상위 정확도를 달성한다. 그러나 이 방법은 네트워크가 복잡하기 때문에 학습과 테스트에 시간이 오래 걸린다는 단점이 있다.

[0161] 3개의 데이터셋에 대한 실험에서 일반적으로 NeT와 기본 ViT를 결합하면 성능이 크게 향상될 수 있다는 것을 발견했다. Swin 변환기가 NeT의 백본으로 사용되었을 때 전체적인 성능은 높았지만, Stanford Dogs 데이터셋에서는 약간 낮은 성능을 보였다.

[0163] 애플리케이션 연구 1 - 트리 깊이의 효과

[0164] 트리 구조를 정의하는 핵심 요소인 트리 깊이에 따른 성능 변화를 검증하기 위해 CUB-200-2011, Stanford Dogs 및 Stanford Cars 데이터셋에 서로 다른 트리 깊이 {3, 4, 5, 6, 7}의 다섯 가지를 정의했다.

[0166] 도 8은 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 트리 깊이에 따른 정확도를 나타낸 도면이다. 도 8에 도시된 바와 같이, ViT-NeT는 CUB-200-2011 데이터셋에서 트리의 깊이를 4로 설정할 때 최고의 성능을 달성한다. 깊이가 3인 경우 모델은 제한된 용량을 가지며, 이는 데이터셋의 미묘한 차이를 나타내기에는 불충분하다.

- [0168] 반면에 트리의 깊이를 5 이상으로 설정하면, 데이터셋에 대한 과도한 수의 학습 가능한 매개 변수가 모델을 과적합 상태로 이끌게 된다. 도 9는 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 트리 깊이가 5일 때 특정 의사결정 경로에서의 프로토타입 반응을 나타낸 도면이다. 도 9에 도시된 바와 같이, 트리의 깊이가 증가함에 따라 모델 과적합이 발생해 새의 날개 끝에 동일한 프로토타입 반응이 나타난다. 이 문제는 의사결정 프로세스의 순차적 분석을 통해서도 검증될 수 있으며, 활성화가 주어진 이미지의 중복 특성에서 과생된다는 것을 확인할 수 있다.
- [0170] 한편, 도 8에 도시된 바와 같이, Stanford Cars 및 Stanford Dogs 데이터셋의 경우, 트리 깊이가 5 이상일 때 상당한 성능을 보였다. Stanford Cars 데이터셋의 경우 트리 깊이에 따라 성능에서 큰 차이가 없었으며, 최대 성능은 깊이 6에서 발생했다. 마찬가지로, Stanford Dogs 데이터셋에서 최고의 성능은 6의 깊이에서 달성되었지만, 트리 깊이에 따라 성능 차이가 더 컸다. 실험 결과에서 NeT의 트리 깊이는 데이터셋의 복잡성에 따라 달라야 하며, 깊이를 높이면 성능이 향상되는 것이 아니라 오히려 과적합으로 이어진다는 것을 알 수 있다.
- [0172] 애플리케이션 연구 2 - 리프 노드의 유효성
- [0173] 리프 노드는 라우팅된 패치를 서브 카테고리 공간에 투영하는 데 중요한 역할을 한다. 본 발명에서는 리프 노드가 강인한 집약된 패치를 학습할 수 있도록 GAP와 GMP의 두 가지 옵션을 고려했다.
- [0175] 도 10은 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법의 풀링 방법에 따른 실험 결과를 나타낸 도면이다. 도 10에 제시된 비교 결과, GMP를 사용하면 CUB-200-2011 데이터셋에서 GAP를 제치고 1위 정확도(0.2%)를 얻을 수 있었다.
- [0177] 애플리케이션 연구 3 - 맥락 변환기 모듈의 효과
- [0178] 본 발명에서는 모든 트리 에지에서 CTM을 사용하여 뉴럴 트리 디코더(200)를 설계하였다. CTM의 효과를 평가하기 위해 2개의 백본 네트워크에서 CTM의 유무에 따른 성능 비교를 실시했다.
- [0180] 도 11은 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 맥락 변환기 모듈에 따른 실험 결과를 나타낸 도면이다. 도 11에 도시된 바와 같이, DeiT를 ViT-NeT의 CTM과 함께 백본으로 사용할 경우 상위 1의 정확도가 89.3%에서 90.1%로 높아진다. 마찬가지로 CTM과 함께 SwinT를 백본으로 사용할 경우 상위 1의 정확도는 90.7%에서 91.6%로 높아진다. CTM이 압축된 이미지 패치 내에서 다양한 유형의 전역 맥락 정보를 통합하여 더 나은 성능을 달성할 수 있음을 확인할 수 있다.
- [0182] 정성 분석
- [0183] 본 발명에서는 본질적으로 해석 가능한 NeT 방법을 제시한다. 의사결정의 해석 가능성을 검증하기 위해 무작위로 샘플링된 이미지에 대해 시각화된 순차적 의사결정을 보여주는 정성적 평가를 제공한다.
- [0185] 도 12는 본 발명의 일실시예에 따른 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에서, 무작위로 샘플링된 이미지에 대한 순차적 의사결정을 시각화해 도시한 도면이다. 도 12에 도시된 바와 같이, 본 발명의 ViT-NeT는 인간 친화적인 방식으로 지역 차별적 부분을 찾아낸다.
- [0187] 전술한 바와 같이, 본 발명에서 제안하고 있는 뉴럴 트리 디코더(200)를 이용한 해석 가능한 비전 변환기 및 그에 의한 분류 방법에 따르면, 비전 변환기(ViT) 인코더와 뉴럴 트리(NeT) 디코더를 결합해 ViT-NeT를 구성하되,

비전 변환기를 백본으로 사용해 주의 맵을 추출하고 이를 뉴럴 트리 디코더(200)에 입력함으로써, 클래스 간 상관관계가 높고 클래스 내 상관관계가 낮은 세분화된 개체를 정확하게 분류할 수 있고, 트리 구조와 프로토타입을 통해 의사결정 과정을 설명하고 결과를 시각적으로 해석할 수 있어, 기존 앙상블 모델에서 발생하는 해석 가능성과 성능의 트레이드 오프를 해결할 수 있다.

[0189] 한편, 본 발명은 다양한 통신 단말기로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터에서 판독 가능한 매체를 포함할 수 있다. 예를 들어, 컴퓨터에서 판독 가능한 매체는, 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) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다.

[0191] 이와 같은 컴퓨터에서 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이때, 컴퓨터에서 판독 가능한 매체에 기록되는 프로그램 명령은 본 발명을 구현하기 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 예를 들어, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

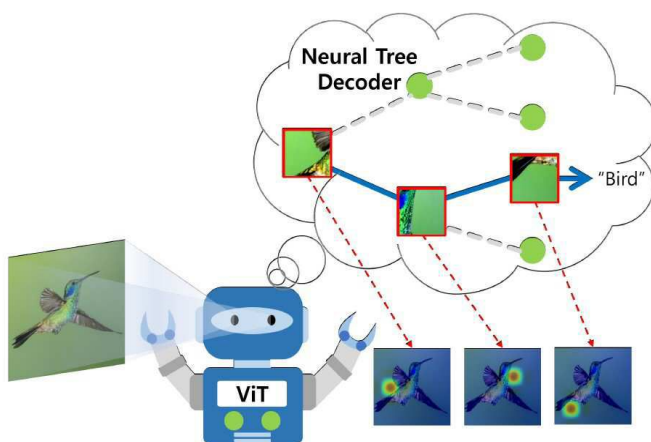
[0193] 이상 설명한 본 발명은 본 발명이 속한 기술분야에서 통상의 지식을 가진 자에 의하여 다양한 변형이나 응용이 가능하며, 본 발명에 따른 기술적 사상의 범위는 아래의 특허청구범위에 의하여 정해져야 할 것이다.

부호의 설명

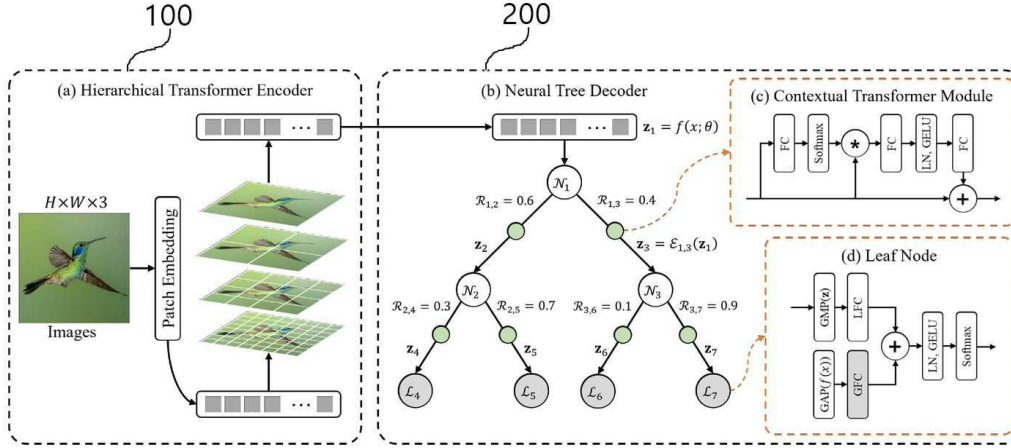
[0194] 100: 계층적 변환기 인코더
200: 뉴럴 트리 디코더
S100: 계층적 변환기 인코더 및 뉴럴 트리 디코더를 포함하여 구성되는 ViT-NeT를 학습해 분류 모델을 생성하는 단계
S200: 분류 모델에 입력 이미지를 입력해 최종 클래스 레이블을 예측하는 단계

도면

도면1



도면2



도면3

Algorithm 1 Training a ViT-NeT

Input: Training set $\mathcal{T}$, number of epochs E , hierarchical transformer encoder f , depth of tree d , encoder parameter θ , decoder parameter ω

Initialize $\mathbb{T} = \text{init tree}(d; \omega)$

Initialize leaf node set $\mathbb{L} = \emptyset$

Get leaf node ids from $\mathbb{T}$ and assign it into $\mathbb{L}$

function $\text{traverse}(i, z_i)$

if $i \in \mathbb{L}$ **then**

return $\mathcal{L}(z_i)$

end if

$\mathcal{R}_{i, 2 \times i}(z_i) = [\mathcal{N}(z_i)]_0^1$

$\mathcal{R}_{i, 2 \times i + 1}(z_i) = 1 - [\mathcal{N}(z_i)]_0^1$

$l_{\text{distrib}} = \text{traverse}(2 \times i, \mathcal{E}_{i, 2 \times i}(z_i))$

$r_{\text{distrib}} = \text{traverse}(2 \times i + 1, \mathcal{E}_{i, 2 \times i + 1}(z_i))$

return $\mathcal{R}_{i, 2 \times i}(z_i) \times l_{\text{distrib}} + \mathcal{R}_{i, 2 \times i + 1}(z_i) \times r_{\text{distrib}}$

end function

for $e \in \{1, \dots, E\}$ **do**

 randomly split $\mathcal{T}$ into B mini-batches

for $(x_b, y_b) \in \{\mathcal{T}_1, \dots, \mathcal{T}_b, \dots, \mathcal{T}_B\}$ **do**

$z_1^b = f(x_b; \theta)$

$\hat{y}_b = \text{traverse}(1, z_1^b)$

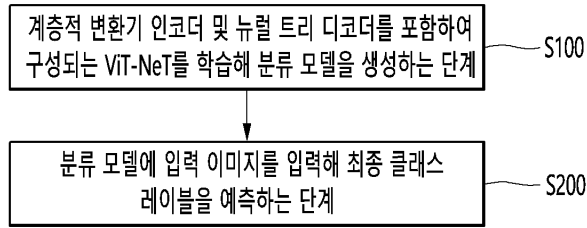
 compute $\text{loss}(\hat{y}_b, y_b)$

 update parameter θ and ω with AdamW

end for

end for

도면4



도면5

Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	82.2
STN (Jaderberg et al., 2015)	Inception	84.1
ResNet-50 (He et al., 2016)	ResNet-50	84.5
MA-CNN (Zheng et al., 2017)	VGG-19	86.5
DCL (Chen et al., 2019b)	VGG-16	86.9
TASN (Zheng et al., 2019b)	VGG-19	87.1
DFL-CNN (Wang et al., 2018)	ResNet-50	87.4
NTS-Net (Yang et al., 2018)	ResNet-101	87.9
DCL (Chen et al., 2019b)	ResNet-50	87.8
TASN (Zheng et al., 2019b)	ResNet-50	87.9
DBTNet (Zheng et al., 2019a)	ResNet-101	88.1
FDL (Liu et al., 2020)	DenseNet-161	89.1
PMG (Du et al., 2020)	ResNet-50	89.6
API-Net (Zhuang et al., 2020)	DenseNet-161	90.0
StackedLSTM (Ge et al., 2019)	GoogleNet	90.4
DeiT (Touvron et al., 2021)	DeiT-B	87.6
SwinT (Liu et al., 2021)	SwinT-B	88.4
TransFG (He et al., 2021)	ViT-B/16	90.9
AFTrans (Zhang et al., 2021)	ViT-B/16	91.5
ViT-NeT	DeiT-B	90.1
ViT-NeT	SwinT-B	91.6

† 224 image input

도면6

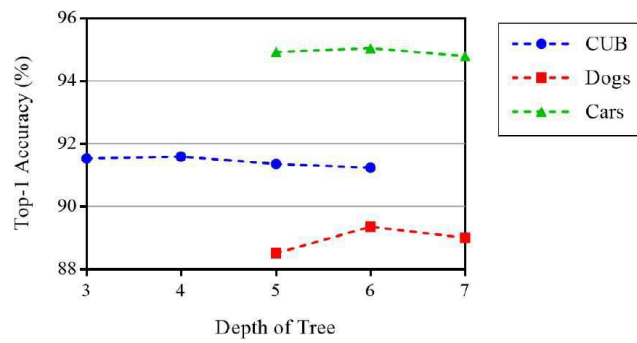
Method	Backbone	Top-1 (%)
MaxEnt (Dubey et al., 2018)	DenseNet-161	83.6
FDL (Liu et al., 2020)	DenseNet-161	84.9
DFL-CNN (Wang et al., 2018)	ResNet-50	84.9
RA-CNN (Fu et al., 2017)	VGG-19	87.3
Cross-X (Luo et al., 2019)	ResNet-50	88.9
SEF (Luo et al., 2020)	ResNet-50	88.8
API-Net (Zhuang et al., 2020)	ResNet-101	90.3
DeiT (Touvron et al., 2021)	DeiT-B	91.5
SwinT (Liu et al., 2021)	SwinT-B	88.0
TransFG (He et al., 2021)	ViT-B/16	90.4
AFTrans (Zhang et al., 2021)	ViT-B/16	91.6
ViT-NeT	DeiT-B	93.6
ViT-NeT	SwinT-B	90.3

도면7

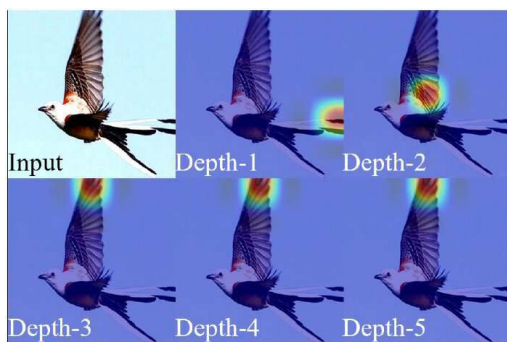
Method	Backbone	Top-1 (%)
ProtoTree† (Nauta et al., 2021)	ResNet-50	86.6
RA-CNN (Fu et al., 2017)	VGG-19	92.5
MaxEnt (Dubey et al., 2018)	DenseNet-161	93.0
DFL-CNN (Wang et al., 2018)	ResNet-50	93.1
SEF (Luo et al., 2020)	ResNet-50	94.0
FDL (Liu et al., 2020)	DenseNet-161	94.2
Cross-X (Luo et al., 2019)	ResNet-50	94.6
MMAL (Balikas et al., 2017)	ResNet-50	95.0
PMG (Du et al., 2020)	ResNet-50	95.1
API-Net (Zhuang et al., 2020)	DenseNet-161	95.3
DeiT (Touvron et al., 2021)	DeiT-B	92.4
SwinT (Liu et al., 2021)	SwinT-B	94.5
TransFG (He et al., 2021)	ViT-B/16	94.1
AFTrans (Zhang et al., 2021)	ViT-B/16	95.0
ViT-NeT	DeiT-B	94.7
ViT-NeT	SwinT-B	<u>95.0</u>

† 224 image input

도면8



도면9



도면10

Pooling	Top-1 (%)
GAP	91.4
GMP	91.6

도면11

Backbone	Use CTM	Top-1 (%)
DeiT-B	✗	89.3
	✓	90.1
SwinT-B	✗	90.7
	✓	91.6

도면12

