

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0146400

(43) 공개일자 2023년10월19일

(51) 국제특허분류(Int. Cl.)

G06V 10/764 (2022.01) G06V 10/46 (2022.01)

G06V 10/70 (2022.01) G06V 10/774 (2022.01)

G06V 10/82 (2022.01)

(52) CPC특허분류

G06V 10/764 (2023.08)

G06V 10/467 (2023.08)

(21) 출원번호 10-2022-0045401

(22) 출원일자 2022년04월12일

심사청구일자 2022년04월12일

(71) 출원인

계명대학교 산학협력단

대구광역시 달서구 달구벌대로 1095, 계명대학교
산학협력관 201호(신당동)

(72) 발명자

고병철

대구광역시 수성구 범어천로 107 범어 STX KAN,
101동 1002호

안다습

대구광역시 북구 침산로21길 39 침산태왕아너스르
템, 603호

김상원

대구광역시 수성구 지범로22길 20 지산현대맨션,
105동 905호

(74) 대리인

김건우

전체 청구항 수 : 총 16 항

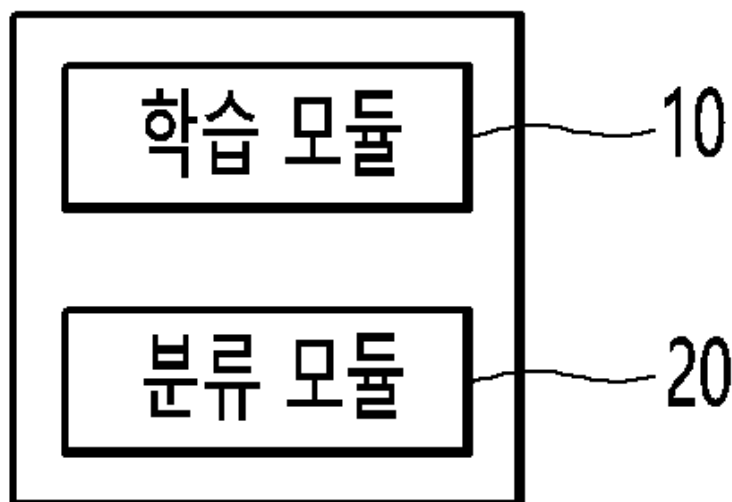
(54) 발명의 명칭 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치

(57) 요약

본 발명은 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법에 관한 것으로서, 보다 구체적으로는 컴퓨터에 의해 각 단계가 수행되는 이미지 분류 방법으로서, (1) 학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기를 학습하는 단계; 및 (2) 상기 학습된 삼 비전 변환기를

(뒷면에 계속)

대표도 - 도2



이용해 이미지를 분류하는 단계를 포함하며, 상기 단계 (1)에서는, 상기 학습 데이터에 포함되는 입력 이미지와 상기 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 상기 삼 비전 변환기를 학습하는 것을 그 구성상의 특징으로 한다.

또한, 본 발명은 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 장치에 관한 것으로서, 보다 구체적으로는 이미지 분류 장치로서, 학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기를 학습하는 학습 모듈; 및 상기 학습된 삼 비전 변환기를 이용해 이미지를 분류하는 분류 모듈을 포함하며, 상기 학습 모듈은, 상기 학습 데이터에 포함되는 입력 이미지와 상기 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 상기 삼 비전 변환기를 학습하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에 따르면, 입력 이미지의 수직 벡터값과 수평 벡터값을 삼 구조의 ViT 인코더에 각각 입력해 삼 비전 변환기를 학습함으로써, 기존 비전 변환기보다 이미지 분류 성능이 우수하다.

(52) CPC특허분류

G06V 10/469 (2023.08)

G06V 10/774 (2023.08)

G06V 10/82 (2022.01)

G06V 10/87 (2023.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345344255
과제번호	2019R1I1A3A01042506
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	지역대학우수과학자지원사업(1년~5년)
연구과제명	역전과가 필요 없는 심층 랜덤 네트워크 모델 개발과 응용 연구
기 여 율	1/1
과제수행기관명	계명대학교
연구기간	2022.03.01 ~ 2022.05.31

명세서

청구범위

청구항 1

컴퓨터에 의해 각 단계가 수행되는 이미지 분류 방법으로서,

(1) 학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기(100)를 학습하는 단계; 및

(2) 상기 학습된 삼 비전 변환기(100)를 이용해 이미지를 분류하는 단계를 포함하며,

상기 단계 (1)에서는,

상기 학습 데이터에 포함되는 입력 이미지와 상기 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 상기 삼 비전 변환기(100)를 학습하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 2

제1항에 있어서, 상기 삼 비전 변환기(100)는,

상기 입력 이미지에서 수평 벡터값과 수직 벡터값을 생성하는 입력 벡터 생성부(110); 및

메모리와 가중치를 공유하는 삼 구조의 ViT 인코더로 구성되는 삼 인코더부(120)를 포함하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 3

제2항에 있어서, 상기 입력 벡터 생성부(110)는,

상기 입력 이미지를 복수의 패치로 나누어 정렬한 수평 벡터값과, 상기 입력 이미지를 시프트(shift)한 변이 이미지를 복수의 패치로 나누어 정렬한 수직 벡터값을 각각 생성하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 4

제3항에 있어서, 상기 입력 벡터 생성부(110)는,

상기 입력 이미지 및 변이 이미지에 위치값과 클래스 토큰값을 추가하는 패치 임베딩을 수행해 상기 수평 벡터값과 수직 벡터값을 생산하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 5

제2항에 있어서, 상기 삼 구조의 ViT 인코더는,

각각 복수의 트랜스포머 레이어를 포함하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 6

제5항에 있어서, 상기 트랜스포머 레이어는,

2개의 노름(Norm), 멀티 헤드 어텐션(MHA) 및 다층 퍼셉트론(MLP)을 포함하여 구성되는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 7

제5항에 있어서, 상기 삼 인코더부(120)는,

상기 수평 벡터값과 수직 벡터값을 상기 삼 구조의 ViT 인코더를 통해 학습하며, 마지막 트랜스포머 레이어의 클래스 토큰을 다층 퍼셉트론에 적용해 분류 확률값을 도출하고, 상기 삼 구조의 ViT 인코더에서 각각 도출된 분류 확률값을 결합해 최종 분류를 하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 8

제1항에 있어서, 상기 단계 (1)에서는,

교차 엔트로피(Cross Entropy) 기반의 손실함수를 이용해 상기 삼 비전 변환기(100)를 학습하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법.

청구항 9

이미지 분류 장치로서,

학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기(100)를 학습하는 학습 모듈(10); 및

상기 학습된 삼 비전 변환기(100)를 이용해 이미지를 분류하는 분류 모듈(20)을 포함하며,

상기 학습 모듈(10)은,

상기 학습 데이터에 포함되는 입력 이미지와 상기 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 상기 삼 비전 변환기(100)를 학습하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 10

제9항에 있어서, 상기 삼 비전 변환기(100)는,

상기 입력 이미지에서 수평 벡터값과 수직 벡터값을 생성하는 입력 벡터 생성부(110); 및

메모리와 가중치를 공유하는 삼 구조의 ViT 인코더로 구성되는 삼 인코더부(120)를 포함하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 11

제10항에 있어서, 상기 입력 벡터 생성부(110)는,

상기 입력 이미지를 복수의 패치로 나누어 정렬한 수평 벡터값과, 상기 입력 이미지를 시프트(shift)한 변이 이미지를 복수의 패치로 나누어 정렬한 수직 벡터값을 각각 생성하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 12

제11항에 있어서, 상기 입력 벡터 생성부(110)는,

상기 입력 이미지 및 변이 이미지에 위칫값과 클래스 토큰값을 추가하는 패치 임베딩을 수행해 상기 수평 벡터값과 수직 벡터값을 생산하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 13

제10항에 있어서, 상기 삼 구조의 ViT 인코더는,

각각 복수의 트랜스포머 레이어를 포함하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 14

제13항에 있어서, 상기 트랜스포머 레이어는,

2개의 노름(Norm), 멀티 헤드 어텐션(MHA) 및 다층 퍼셉트론(MLP)을 포함하여 구성되는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 15

제13항에 있어서, 상기 삼 인코더부(120)는,

상기 수평 벡터값과 수직 벡터값을 상기 삼 구조의 ViT 인코더를 통해 학습하며, 마지막 트랜스포머 레이어의 클래스 토큰을 다층 퍼셉트론에 적용해 분류 확률값을 도출하고, 상기 삼 구조의 ViT 인코더에서 각각 도출된 분류 확률값을 결합해 최종 분류를 하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

청구항 16

제9항에 있어서, 상기 학습 모듈(10)은,

교차 엔트로피(Cross Entropy) 기반의 손실함수를 이용해 상기 삼 비전 변환기(100)를 학습하는 것을 특징으로 하는, 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치.

발명의 설명

기술 분야

[0001] 본 발명은 이미지 분류 방법 및 장치에 관한 것으로서, 보다 구체적으로는 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에 관한 것이다.

배경 기술

[0002] 변환기(Transformer, 트랜스포머)는 현재 자연어처리와 같은 언어적 분야에서 우수한 성능으로 크게 주목받고 있는 기술이다. 현재는 변환기의 학습 메커니즘을 비전 도메인에 적용하기 위한 비전 변환기(Vision Transformer, ViT) 연구가 활발히 진행되고 있다. 비전 변환기는 셀프 어텐션 메커니즘을 사용해 성능이 우수할 뿐만 아니라, 계산 효율성과 확장성이 좋아 메모리를 효율적으로 사용하며 연산 속도 또한 개선할 수 있다. 이러한 장점으로 인해, 최근에는 CNN 기반의 영상 이미지 분류보다 비전 변환기 기반의 영상 이미지 분류 연구가 더 활발히 이루어지고 있다.

[0004] 비전 변환기는 이미지의 서로 다른 부분 간의 의존성을 모델링하여 전체적인 관계를 확인하는 데 적합하기 때문에 이미지의 전역적인 특징 학습에 강점을 보인다. 하지만 다양한 크기의 객체와 객체의 방향 변화에 대해서는 예측하기 힘들기 때문에 방대한 양의 다양한 데이터가 필요하다. 즉, 비전 변환기는 기존의 CNN 모델과 달리 귀납적 편향이 부족하므로 영상의 작은 변화에도 민감하게 반응하는 한계가 있다. 따라서 이러한 객체의 귀납적 편향을 예측할 수 있는 새로운 비전 변환기 알고리즘의 개발이 필요하다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 입력 이미지의 수직 벡터값과 수평 벡터값을 삼 구조의 ViT 인코더에 각각 입력해 삼 비전 변환기를 학습함으로써, 기존 비전 변환기보다 이미지 분류 성능이 우수한, 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치를 제공하는 것을 그 목적으로 한다.

과제의 해결 수단

[0006] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법은,

[0007] 컴퓨터에 의해 각 단계가 수행되는 이미지 분류 방법으로서,

[0008] (1) 학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기를 학습하는 단계; 및

[0009] (2) 상기 학습된 삼 비전 변환기를 이용해 이미지를 분류하는 단계를 포함하며,

[0010] 상기 단계 (1)에서는,

[0011] 상기 학습 데이터에 포함되는 입력 이미지와 상기 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 상기 삼 비전 변환기를 학습하는 것을 그 구성상의 특징으로 한다.

[0013] 바람직하게는, 상기 삼 비전 변환기는,

[0014] 상기 입력 이미지에서 수평 벡터값과 수직 벡터값을 생성하는 입력 벡터 생성부; 및

[0015] 메모리와 가중치를 공유하는 삼 구조의 ViT 인코더로 구성되는 삼 인코더부를 포함할 수 있다.

[0017] 더욱 바람직하게는, 상기 입력 벡터 생성부는,

[0018] 상기 입력 이미지를 복수의 패치로 나누어 정렬한 수평 벡터값과, 상기 입력 이미지를 시프트(shift)한 변이 이미지를 복수의 패치로 나누어 정렬한 수직 벡터값을 각각 생성할 수 있다.

[0020] 더더욱 바람직하게는, 상기 입력 벡터 생성부는,

[0021] 상기 입력 이미지 및 변이 이미지에 위치값과 클래스 토큰값을 추가하는 패치 임베딩을 수행해 상기 수평 벡터값과 수직 벡터값을 생산할 수 있다.

[0023] 더욱 바람직하게는, 상기 삼 구조의 ViT 인코더는,

[0024] 각각 복수의 트랜스포머 레이어를 포함할 수 있다.

- [0026] 더더욱 바람직하게는, 상기 트랜스포머 레이어는,
- [0027] 2개의 노름(Norm), 멀티 헤드 어텐션(MHA) 및 다층 퍼셉트론(MLP)을 포함하여 구성될 수 있다.
- [0029] 더더욱 바람직하게는, 상기 삼 인코더부는,
- [0030] 상기 수평 벡터값과 수직 벡터값을 상기 삼 구조의 ViT 인코더를 통해 학습하며, 마지막 트랜스포머 레이어의 클래스 토큰을 다층 퍼셉트론에 적용해 분류 확률값을 도출하고, 상기 삼 구조의 ViT 인코더에서 각각 도출된 분류 확률값을 결합해 최종 분류를 할 수 있다.
- [0032] 바람직하게는, 상기 단계 (1)에서는,
- [0033] 교차 엔트로피(Cross Entropy) 기반의 손실함수를 이용해 상기 삼 비전 변환기를 학습할 수 있다.
- [0035] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 장치는,
- [0036] 이미지 분류 장치로서,
- [0037] 학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기를 학습하는 학습 모듈; 및
- [0038] 상기 학습된 삼 비전 변환기를 이용해 이미지를 분류하는 분류 모듈을 포함하며,
- [0039] 상기 학습 모듈은,
- [0040] 상기 학습 데이터에 포함되는 입력 이미지와 상기 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 상기 삼 비전 변환기를 학습하는 것을 그 구성상의 특징으로 한다.
- [0042] 바람직하게는, 상기 삼 비전 변환기는,
- [0043] 상기 입력 이미지에서 수평 벡터값과 수직 벡터값을 생성하는 입력 벡터 생성부; 및
- [0044] 메모리와 가중치를 공유하는 삼 구조의 ViT 인코더로 구성되는 삼 인코더부를 포함할 수 있다.
- [0046] 더욱 바람직하게는, 상기 입력 벡터 생성부는,
- [0047] 상기 입력 이미지를 복수의 패치로 나누어 정렬한 수평 벡터값과, 상기 입력 이미지를 시프트(shift)한 변이 이미지를 복수의 패치로 나누어 정렬한 수직 벡터값을 각각 생성할 수 있다.
- [0049] 더더욱 바람직하게는, 상기 입력 벡터 생성부는,
- [0050] 상기 입력 이미지 및 변이 이미지에 위치값과 클래스 토큰값을 추가하는 패치 임베딩을 수행해 상기 수평 벡터값과 수직 벡터값을 생산할 수 있다.
- [0052] 더욱 바람직하게는, 상기 삼 구조의 ViT 인코더는,
- [0053] 각각 복수의 트랜스포머 레이어를 포함할 수 있다.
- [0055] 더더욱 바람직하게는, 상기 트랜스포머 레이어는,

[0056] 2개의 노름(Norm), 멀티 헤드 어텐션(MHA) 및 다층 퍼셉트론(MLP)을 포함하여 구성될 수 있다.

[0058] 더더욱 바람직하게는, 상기 삼 인코더부는,

[0059] 상기 수평 벡터값과 수직 벡터값을 상기 삼 구조의 ViT 인코더를 통해 학습하며, 마지막 트랜스포머 레이어의 클래스 토큰을 다층 퍼셉트론에 적용해 분류 확률값을 도출하고, 상기 삼 구조의 ViT 인코더에서 각각 도출된 분류 확률값을 결합해 최종 분류를 할 수 있다.

[0061] 바람직하게는, 상기 학습 모듈은,

[0062] 교차 엔트로피(Cross Entropy) 기반의 손실함수를 이용해 상기 삼 비전 변환기를 학습할 수 있다.

발명의 효과

[0063] 본 발명에서 제안하고 있는 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에 따르면, 입력 이미지의 수직 벡터값과 수평 벡터값을 삼 구조의 ViT 인코더에 각각 입력해 삼 비전 변환기를 학습함으로써, 기존 비전 변환기보다 이미지 분류 성능이 우수하다.

도면의 간단한 설명

[0064] 도 1은 비전 변환기의 전체 구조를 도시한 도면.

도 2는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 장치의 구성을 도시한 도면.

도 3은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법의 흐름을 도시한 도면.

도 4는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기의 구성을 도시한 도면.

도 5는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기의 전체 구조를 도시한 도면.

도 6은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기의 전체 구조를 블록으로 나타낸 도면.

도 7은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에서, 벡터값 추출 과정을 예를 들어 도시한 도면.

도 8은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에서, 벡터값을 삼 구조의 인코더에 입력해 학습하는 과정을 도시한 도면.

도 9는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에서, 삼 구조 인코더를 도시한 도면.

도 10은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치와 최신 변환기 기반 방법의 성능을 Cifar100 데이터세트에 대해 비교한 결과를 나타낸 도면.

발명을 실시하기 위한 구체적인 내용

[0065] 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.

- [0067] 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.
- [0069] 본 발명은 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치에 관한 것으로서, 본 발명의 특징에 따른 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치는, 메모리 및 프로세서를 포함한 하드웨어에서 기록되는 소프트웨어로 구성될 수 있다. 예를 들어, 본 발명의 변이 브랜치를 사용한 삼 비전 변환기에 의한 이미지 분류 방법 및 장치는, 개인용 컴퓨터, 노트북 컴퓨터, 서버 컴퓨터, PDA, 스마트폰, 태블릿 PC 등에 저장 및 구현될 수 있다. 이하에서는 설명의 편의를 위해, 각 단계를 수행하는 주체는 생략될 수 있다.
- [0071] 도 1은 비전 변환기의 전체 구조를 도시한 도면이다. 도 1에 도시된 바와 같이, 일반 비전 변환기는 원본 이미지를 정해진 패치 크기로 나누어 일렬로 정렬한 벡터값을 인코더 입력 값으로 사용한다. 본 발명에서는, 이와 같은 비전 변환기의 구조를 변경해, 변이 브랜치를 사용한 삼 비전 변환기(Shift-ViT)를 제안하고자 한다.
- [0073] 도 2는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치의 구성을 도시한 도면이다. 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 장치는, 학습 데이터를 이용해, 비전 변환기(Vision Transformer, ViT)와 삼 구조를 결합한 삼 비전 변환기(100)(Seiamese Vision Transformer using Shifted Branches: Shift-ViT)를 학습하는 학습 모듈(10); 및 학습된 삼 비전 변환기(100)를 이용해 이미지를 분류하는 분류 모듈(20)을 포함하여 구성될 수 있으며, 학습 모듈(10)은, 학습 데이터에 포함되는 입력 이미지와 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 삼 비전 변환기(100)(Shift-ViT)를 학습할 수 있다.
- [0075] 도 3은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법의 흐름을 도시한 도면이다. 도 3에 도시된 바와 같이, 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법은, 컴퓨터에 의해 각 단계가 수행되는 이미지 분류 방법으로서, 학습 데이터를 이용해 삼 비전 변환기(100)를 학습하는 단계(S10) 및 학습된 삼 비전 변환기(100)를 이용해 이미지를 분류하는 단계(S20)를 포함하여 구현될 수 있으며, 단계 S10에서는, 학습 데이터에 포함되는 입력 이미지와 입력 이미지를 시프트(shift)한 변이 이미지를 삼 구조의 ViT 인코더에 각각 입력해 삼 비전 변환기(100)를 학습할 수 있다.
- [0077] 도 4는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기(100)의 구성을 도시한 도면이고, 도 5는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기(100)의 전체 구조를 도시한 도면이고, 도 6은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기(100)의 전체 구조를 블록으로 나타낸 도면이다. 도 4 내지 도 6에 도시된 바와 같이, 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)는, 입력 벡터 생성부(110) 및 삼 인코더부(120)를 포함하여 구성될 수 있다.
- [0079] 즉, 비전 변환기의 인코더 입력 벡터값을 수직 벡터값과 수평 벡터값을 가지도록 분할하고, 이러한 두 개의 입력 벡터를 처리할 수 있도록 삼 구조의 인코더를 사용해 메모리적 효율성을 가지면서 성능을 향상시킬 수 있다.
- [0081] 도 5에 도시된 바와 같이, 삼 비전 변환기(100)(Shift-ViT)의 입력 벡터 생성부(110)는, 도 1에 도시된 바와 같은 기존 비전 변환기와 마찬가지로 224x224 크기의 이미지를 입력으로 받으며, 입력 이미지를 16x16 크기의 패

치로 나누고 위치값과 클래스 토큰값을 추가하는 Patch Embedding을 하게 된다. 이때, 삼 비전 변환기(100)(Shift-ViT)는 기존 비전 변환기와 달리, 입력값으로 수직 벡터값과 수평 벡터값을 만들고, 이후 2개의 벡터값을 삼 인코더부(120)를 구성하는 삼 구조의 인코더 각각의 입력으로 사용할 수 있다. 그 다음, 각각의 간단한 다층 퍼셉트론(MLP)을 사용해 분류 확률값을 도출하고, 2개의 확률값에 대해 평균을 구하여 최종 분류를 할 수 있다.

[0083] 이하에서는, 도 4 내지 도 6을 참조하여, 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 삼 비전 변환기(100)의 각 구성에 대해 상세히 설명하도록 한다.

[0085] 입력 벡터 생성부(110)는, 입력 이미지에서 수평 벡터값과 수직 벡터값을 생성할 수 있다. 보다 구체적으로, 입력 벡터 생성부(110)는, 입력 이미지를 복수의 패치로 나누어 정렬한 수평 벡터값과, 입력 이미지를 시프트(shift)한 변이 이미지를 복수의 패치로 나누어 정렬한 수직 벡터값을 각각 생성할 수 있다. 입력 벡터 생성부(110)는, 입력 이미지 및 변이 이미지에 위치값과 클래스 토큰값을 추가하는 패치 임베딩을 수행해 수평 벡터값과 수직 벡터값을 생산할 수 있다.

[0087] 기존 비전 변환기의 입력 벡터로는 원본 이미지를 16x16 크기의 패치로 나누어 일렬로 정렬한 한 개의 벡터값을 사용한다. 본 발명에서는, 한 개의 벡터값이 아니라, 원본 입력 이미지를 16x16 크기 패치로 나누어 수평으로 패치를 정렬한 벡터값(수평 벡터값)과 수직으로 패치를 정렬한 벡터값(수직 벡터값)을 사용할 수 있도록 Shift 방법을 적용할 수 있다. 이러한 Shifted Branch 방법을 사용하여 이미지의 가로와 세로의 특성을 확인할 수 있게 하며 귀납적 편향을 예측할 수 있다.

[0089] 도 7은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 벡터값 추출 과정을 예를 들어 도시한 도면이다. 도 7에 도시된 바와 같이, 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)의 입력 벡터 생성부(110)는, 원본 이미지(x로 표기)를 16×16 크기 패치로 나누어 정렬한 값과, 원본 이미지를 시프트($\tilde{x}$ 로 표기)하여 16×16 크기 패치로 나누어 정렬한 값을 가지도록 하는 패치 정렬 함수를 사용할 수 있다.

수학식 1

[0091]
$$x \in \mathbb{R}^{224 \times 224 \times c}, \varphi(x) = x_p \in \mathbb{R}^{N \times (16^2 \times c)}$$

수학식 2

[0093]
$$\tilde{x} \in \mathbb{R}^{224 \times 224 \times c}, \varphi(\tilde{x}) = \tilde{x}_p \in \mathbb{R}^{N \times (16^2 \times c)}$$

[0095] 여기서, $\varphi(\cdot)$ 는 Patch Embedding 함수이고, c는 x의 Channel Dimension이다.

[0097] 삼 인코더부(120)는, 메모리와 가중치를 공유하는 삼 구조의 ViT 인코더로 구성될 수 있다. 보다 구체적으로, 삼 구조의 ViT 인코더는, 각각 복수의 트랜스포머 레이어를 포함할 수 있다. 입력 벡터 생성부(110)에서 생성

된 2개의 벡터값은 삼 인코더부(120)가 포함하는 삼 구조로 만들어진 인코더에 입력될 수 있다.

[0099] 도 8은 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 삼 구조 인코더를 도시한 도면이고, 도 9는 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서, 벡터값을 삼 구조의 인코더에 입력해 학습하는 과정을 도시한 도면이다. 도 8 및 도 9에 도시된 바와 같이, 본 발명의 일실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에서는, 메모리와 가중치를 공유하는 효율적인 구조인 삼 구조를 사용할 수 있다. 삼 구조의 인코더는 각각 12개의 트랜스포머 레이어를 가질 수 있으며, 각 트랜스포머 레이어는 2개의 노름(Norm), 멀티 헤드 어텐션(MHA) 및 다층 퍼셉트론(MLP)을 포함하여 구성될 수 있다.

[0101] 삼 인코더부(120)는, 수평 벡터값과 수직 벡터값을 삼 구조의 ViT 인코더를 통해 학습하며, 마지막 트랜스포머 레이어의 클래스 토큰을 간단한 다층 퍼셉트론에 적용해 분류 확률값을 도출하고, 삼 구조의 ViT 인코더에서 각 도출된 분류 확률값을 결합해 최종 분류를 할 수 있다. 삼 인코더부(120)의 학습을 각 패치 z_0 , $\tilde{z}_0$ 에 대하여 다음 수학적식 3 내지 7과 같이 나타낼 수 있다.

수학적식 3

[0103]
$$z_0 = \{x_{cls}, x_p^0, x_p^1, \dots, x_p^N\}$$

수학적식 4

[0105]
$$\tilde{z}_0 = \{x_{cls}, \tilde{x}_p^0, \tilde{x}_p^1, \dots, \tilde{x}_p^N\}$$

수학적식 5

[0107]
$$\begin{aligned} \hat{z}_l &= MHA(LN(z_{l-1})) + z_{l-1}, \quad l = 1, 2 \dots L \\ \tilde{\hat{z}}_l &= MHA(LN(\tilde{z}_{l-1})) + \tilde{z}_{l-1} \end{aligned}$$

수학적식 6

[0109]
$$\begin{aligned} z_l &= MLP(LN(\hat{z}_l)) + \hat{z}_l, \quad l = 1, 2 \dots L \\ \tilde{z}_l &= MLP(LN(\tilde{\hat{z}}_l)) + \tilde{\hat{z}}_l, \end{aligned}$$

수학식 7

$$P = \text{Softmax} (\text{avg} (M L P_e(z_L^0) \oplus M L P_e(\bar{z}_L^0)))$$

여기서, z_0 , $\bar{z}_0$ 는 각각 가로 정렬 패치, 세로 정렬 패치이다. 또한, LN은 Layer Norm을 뜻하고, MHA, MLP는 앞서 언급되었던 멀티 헤드 어텐션(MHA)과 다층 퍼셉트론(MLP)을 뜻한다. $M L P_e$ 는 인코더 결괏값에 대해 확률값을 도출하기 위한 마지막 다층 퍼셉트론(MLP)을 뜻한다.

단계 S10에서는, 교차 엔트로피(Cross Entropy) 기반의 손실함수를 이용해 삼 비전 변환기(100)를 학습할 수 있다. 즉, Shift-ViT 에서는 마지막 결괏값에 대해 Cross Entropy Loss를 사용하여 네트워크를 학습시키며, 손실함수의 공식은 다음 수학식 8과 같다.

수학식 8

$$L = -\sum_{i=1} Y_i \ln(P_i) + (1 - Y_i) \ln(1 - P_i)$$

여기서 Y_i 는 i 번째 이미지의 실측값이다.

실험

전술한 바와 같은 Shift-ViT의 성능 확인을 위해, 비전 변환기 기반의 최신 이미지 분류 알고리즘과 비교하였다. 데이터 세트로는 Cifar100을 사용하였다. Cifar100은 총 100개의 클래스를 포함하며, 클래스당 500개의 훈련 이미지와 100개의 테스트 이미지가 있다.

비전 변환기는 방대한 양의 학습 데이터가 필요하며 그에 따른 사전학습이 중요하다. 따라서 본 실험에서는 ViT-B/16 에 대해 ImageNet21k에 대한 사전학습 가중치를 사용한다. Batch Size는 32, Step은 20000, Optimizer는 SGD를 사용하며, Learning rate는 $3e-2$, Momentum은 0.9로 사용해 실험한다.

도 10은 본 발명의 일 실시예에 따른 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치와 최신 변환기 기반 방법의 성능을 Cifar100 데이터세트에 대해 비교한 결과를 나타낸 도면이다. 본 발명에서 제안하는 알고리즘(Shift-ViT)과 비교하기 위해 변환기 구조가 변형된 알고리즘들과 비교한다. 또한, 비슷한 규모의 알고리즘과 비교하기 위해 각 알고리즘의 기본 버전 성능과 비교한다. 도 10에서 확인할 수 있는 바와 같이, 제안하는 방법(Shift-ViT)은 최첨단 방법 중 가장 성능이 좋은 CeiT-S에 비해 정확도가 약 0.97% 높으며, 기존 비전 변환기(ViT/B-16)보다 0.27% 성능 향상되었다.

전술한 바와 같이, 본 발명에서 제안하고 있는 변이 브랜치를 사용한 삼 비전 변환기(100)에 의한 이미지 분류 방법 및 장치에 따르면, 입력 이미지의 수직 벡터값과 수평 벡터값을 삼 구조의 ViT 인코더에 각각 입력해 삼 비전 변환기(100)를 학습함으로써, 기존 비전 변환기보다 이미지 분류 성능이 우수하다.

[0131] 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.

[0133] 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.

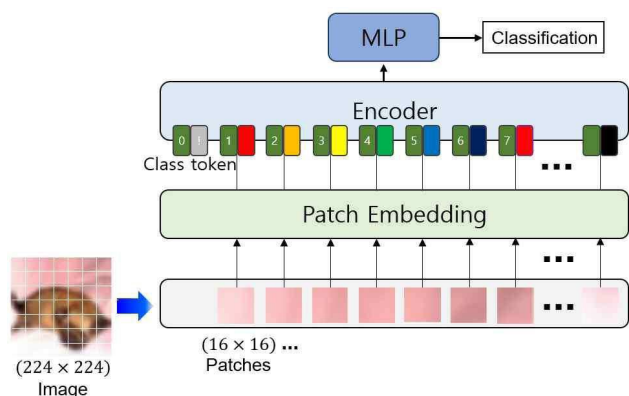
[0135] 이상 설명한 본 발명은 본 발명이 속한 기술분야에서 통상의 지식을 가진 자에 의하여 다양한 변형이나 응용이 가능하며, 본 발명에 따른 기술적 사상의 범위는 아래의 특허청구범위에 의하여 정해져야 할 것이다.

부호의 설명

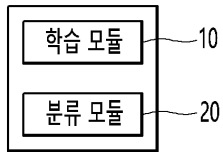
- [0136] 10: 학습 모듈
20: 분류 모듈
100: 삼 비전 변환기
110: 입력 벡터 생성부
120: 삼 인코더부
S10: 학습 데이터를 이용해 삼 비전 변환기를 학습하는 단계
S20: 학습된 삼 비전 변환기를 이용해 이미지를 분류하는 단계

도면

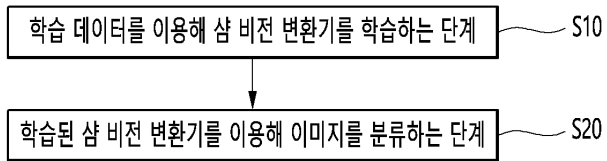
도면1



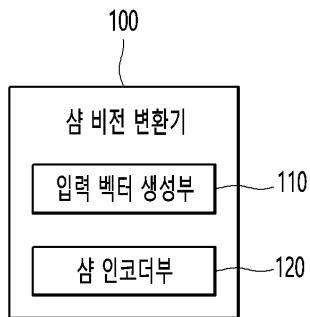
도면2



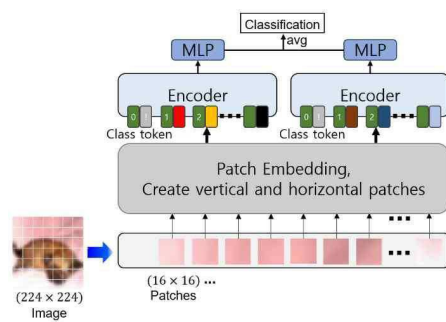
도면3



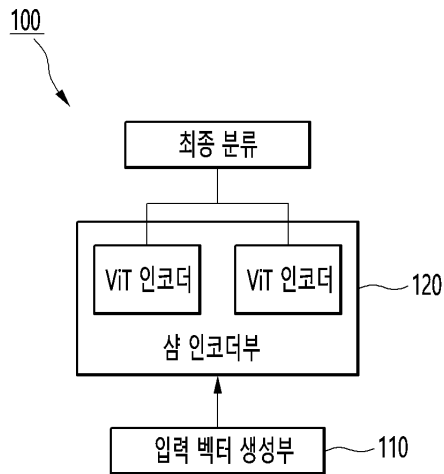
도면4



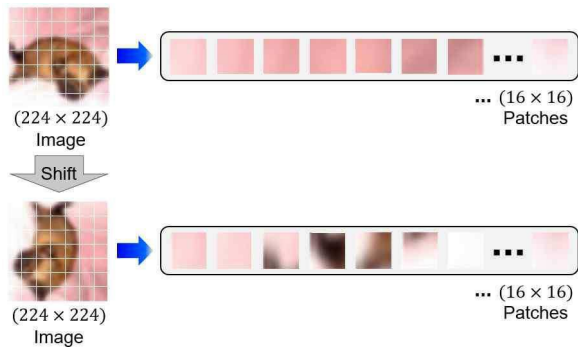
도면5



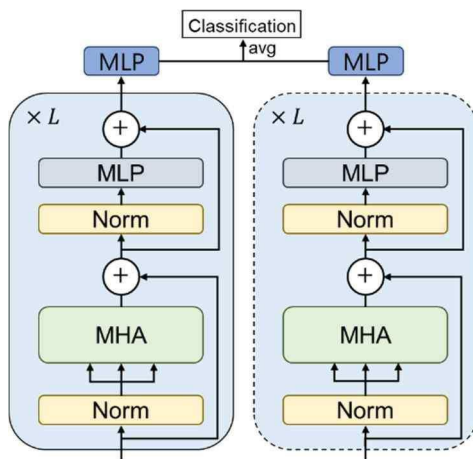
도면6



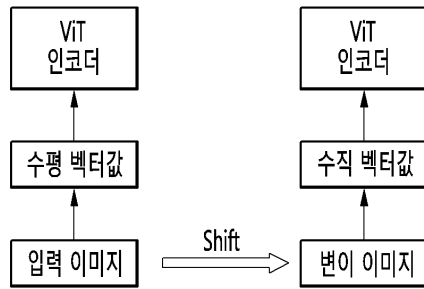
도면7



도면8



도면9



도면10

Method	Accuracy(%)
TNT-B[3]	91.10
DeiT-B[4]	90.80
CeiT-S[5]	91.80
ViT/B-16	92.50
Proposed model (Shift-ViT)	92.77