(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0114893
(43) 공개일자 2023년08월02일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 3/04* (2023.01)
G06V 10/764 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06N 3/084 (2023.01)
G06N 3/047 (2023.01)
(21) 출원번호 10-2022-0011178
(22) 출원일자 2022년01월26일
심사청구일자 2022년01월26일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
양지훈
서울특별시 은평구 백련산로2길 19 백련산힐스테이트1차 101동 404호
정복진
서울특별시 은평구 통일로 590 대림아파트 2동 401호
(74) 대리인
이지연

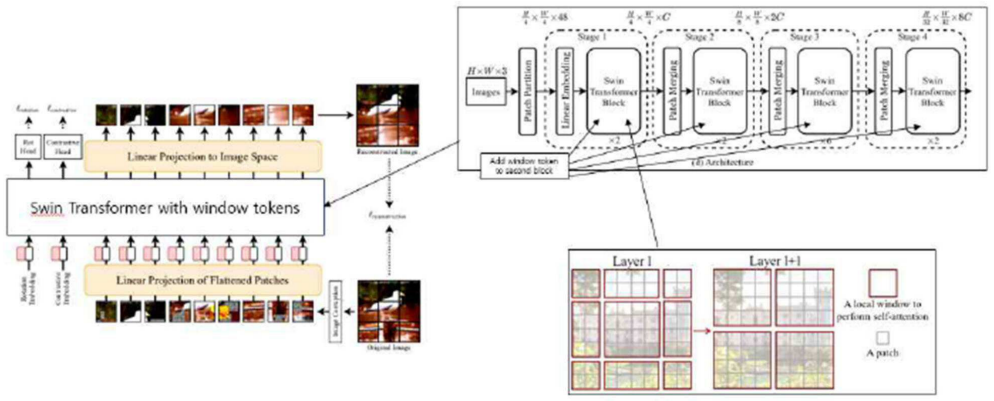
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **자기지도 스윈 트랜스포머 모델 구조 및 이의 학습 방법**

(57) 요 약

본 발명은 자기지도 스윈 트랜스포머 모델 구조에 관한 것이다. 상기 자기지도 스윈 트랜스포머 모델 구조는, 이미지를 입력받는 이미지 입력 모듈; 입력된 이미지를 사전 설정된 크기로 분할하여 패치들을 생성하고, 상기 패치들을 사전 설정된 크기로 묶어 윈도우들을 생성하는 패치 생성 모듈; 각 패치들을 임베딩하여 벡터로 변환하는 패치 임베딩 모듈; 윈도우의 위치를 변환시키는 이동 윈도우 멀티헤드 셀프 어텐션(SW-MSA) 모듈; 로컬 어텐션 알고리즘을 사용하여 어텐션 맵의 크기를 윈도우의 크기에 비례하도록 해주는 윈도우 멀티헤드 셀프 어텐션(W-MAS) 모듈; 인접한 패치들을 하나의 패치로 합쳐주는 패치 머징 모듈;을 구비하고, 상기 SW-MSA 모듈 및 W-MAS 모듈은 쌍으로 이루어져 각 단계별로 반복적으로 실행되고, 마지막 단계에는 하나의 윈도우가 전체 이미지를 감싸는 것을 특징으로 한다.

대 표 도 - 도8



(52) CPC특허분류
G06N 3/049 (2023.01)
G06V 10/764 (2023.01)
G06V 10/82 (2022.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711117075
과제번호	2020-0-00113-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	정보통신방송 연구개발사업
연구과제명	이종 정보 활용 및 데이터 융합을 통한 데이터 증식 기술 개발
기 여 율	1/1
과제수행기관명	서강대학교 산학협력단
연구기간	2020.04.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

이미지를 입력받는 이미지 입력 모듈;

입력된 이미지를 사전 설정된 크기로 분할하여 패치들을 생성하고, 상기 패치들을 사전 설정된 크기로 묶어 윈도우들을 생성하는 패치 생성 모듈;

각 패치들을 임베딩하여 벡터로 변환하는 패치 임베딩 모듈;

윈도우의 위치를 변환시키는 이동 윈도우 멀티헤드 셀프 어텐션(SW-MSA) 모듈;

로컬 어텐션 알고리즘을 사용하여 어텐션 맵의 크기를 윈도우의 크기에 비례하도록 해주는 윈도우 멀티헤드 셀프 어텐션(W-MAS) 모듈;

인접한 패치들을 하나의 패치로 합쳐주는 패치 머징 모듈;

을 구비하고, 상기 SW-MSA 모듈 및 W-MAS 모듈은 쌍으로 이루어져 각 단계별로 반복적으로 실행되고, 마지막 단계에는 하나의 윈도우가 전체 이미지를 감싸는 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조.

청구항 2

제1항에 있어서, 상기 SW-MSA 모듈은

윈도우에서 윈도우 토큰을 제외한 패치들을 사용하여 셀프 어텐션 알고리즘을 수행하여 윈도우 간의 정보를 공유하도록 구성된 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조.

청구항 3

제1항에 있어서, 상기 W-MAS 모듈은,

윈도우 토큰을 패치들과 같이 사용하여 셀프 어텐션 알고리즘을 수행하여 윈도우 토큰에 윈도우의 정보를 집어넣도록 구성된 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조.

청구항 4

제1항에 있어서, 상기 패치 머징 모듈은,

SW-MSA 모듈과 W-MSA 모듈의 쌍이 실행된 후 다음 단계로 넘어갈 때 인접한 4개의 패치를 하나로 만드는 패치 임베딩을 하고, 근접한 윈도우 토큰 4개를 하나로 만드는 토큰 임베딩을 하는 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조.

청구항 5

제1항에 있어서, 상기 패치 생성 모듈은,

각 윈도우에 대해 회전률 토큰과 Contrastive Loss 토큰을 부여하는 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조.

청구항 6

(a) 입력된 이미지를 2개의 이미지로 증가시키는 학습 이미지 증가 단계;

(b) 2개의 증가 이미지를 자기지도 스윈 트랜스포머 모델에 입력하는 단계;

(c) 상기 자기지도 스윈 트랜스포머 모델에서 출력으로 토큰값들을 사용하여 이미지 회전률과 Contrastive Loss를 구하는 단계; 및

(d) 패치값들을 이용하여 본래의 패치로 복원하는 단계;를 구비하고,

상기 자기지도 스윈 트랜스포머 모델은,

이미지를 입력받는 이미지 입력 모듈;

입력된 이미지를 사전 설정된 크기로 분할하여 패치들을 생성하고, 상기 패치들을 사전 설정된 크기로 묶어 윈도우들을 생성하며, 각 윈도우에 회전률 토큰과 CL 토큰을 부여하는 패치 생성 모듈;

각 패치들을 임베딩하여 벡터로 변환하는 패치 임베딩 모듈;

윈도우의 위치를 변환시키는 이동 윈도우 멀티헤드 셀프 어텐션(SW-MSA) 모듈;

로컬 어텐션 알고리즘을 사용하여 어텐션 맵의 크기를 윈도우의 크기에 비례하도록 해주는 윈도우 멀티헤드 셀프 어텐션(W-MAS) 모듈;

인접한 패치들을 하나의 패치로 합쳐주는 패치 머징 모듈;

을 구비하고, 상기 SW-MSA 모듈 및 W-MAS 모듈은 쌍으로 이루어져 각 단계별로 반복적으로 실행되고, 마지막 단계에는 하나의 윈도우가 전체 이미지를 감싸는 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조의 학습 방법.

청구항 7

제6항에 있어서, 상기 (a) 단계는,

입력된 이미지에 대하여 사전 설정된 이미지 증가 기법을 사용하여 2개의 이미지로 증가시키고, 증가된 2개의 이미지를 랜덤하게 회전시킨 후, 임의로 망가뜨리는 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조의 학습 방법.

청구항 8

제6항에 있어서, 상기 SW-MSA 모듈은

윈도우에서 윈도우 토큰을 제외한 패치들을 사용하여 셀프 어텐션 알고리즘을 수행하여 윈도우 간의 정보를 공유하도록 구성된 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조의 학습 방법.

청구항 9

제6항에 있어서, 상기 W-MAS 모듈은,

윈도우 토큰을 패치들과 같이 사용하여 셀프 어텐션 알고리즘을 수행하여 윈도우 토큰에 윈도우의 정보를 집어넣도록 구성된 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조의 학습 방법.

청구항 10

제6항에 있어서, 상기 패치 머징 모듈은,

SW-MSA 모듈과 W-MSA 모듈의 쌍이 실행된 후 다음 단계로 넘어갈 때 인접한 4개의 패치를 하나로 만드는 패치 임베딩을 하고, 근접한 윈도우 토큰 4개를 하나로 만드는 토큰 임베딩을 하는 것을 특징으로 하는 자기지도 스윈 트랜스포머 모델 구조의 학습 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 스윈 트랜스포머 모델 구조 및 이에 대한 자기지도 학습 방법에 관한 것으로서, 더욱 구체적으로는 로컬 어텐션 알고리즘을 적용한 스윈 트랜스포머(Swin Transformer) 모델에 자기지도학습 알고리즘을 적용하고, 윈도우 토큰(Window Token)을 적용한 다중 목적 함수를 활용하여 스윈 트랜스포머 모델을 학습함으로써 비전 트랜스포머의 성능을 향상시키도록 구성된 것을 특징으로 하는 스윈 트랜스포머 모델 구조 및 이에 대한 자기지도 학습 방법에 관한 것이다.

배 경 기 술

[0002] 최근 이미지 분류, 객체 검출, 비디오 분류 등 다양한 컴퓨터 비전 문제에 기계 학습 및 딥러닝을 적용한 모델이 좋은 성능을 보이고 있다. 대부분의 모델은 사람이 직접 정답을 부여하여 레이블된 학습 데이터를 활용하여 지도 학습(Supervised Learning)을 통해 학습시킨다. 이러한 양질의 학습 데이터를 얻기 위하여는 많은 시간과 인력이 필요하다.

[0003] 하지만, 지도 학습을 통해 모델을 학습할 때, 학습 데이터가 적으면 기대 이하의 성능을 내게 된다. 이러한 문제를 해결하기 위하여, 레이블없는 데이터를 활용하여 적은 학습 데이터에도 효율적으로 지도학습을 적용하는 연구가 지속적으로 진행되고 있으며, 그 중 하나가 자기지도학습을 통한 모델 선행 학습 방법이다. 자기지도학습은 스스로 데이터에 임의의 레이블을 붙여준 후 해당 레이블을 사용하여 모델을 학습시키는 방법으로서, 이미 자연어 처리(Natural Language Processing) 분야에서 트랜스포머(Transformer) 기반 모델의 사전 학습에 많이 활용된다. 자기지도학습은 크게 정의된 문제를 해결하면서 모델의 은닉 표현을 찾는 사전 정의 문제(Pretext Task) 기법과 긍정적인 쌍(Positive Pair)과 부정적인 쌍(Negative Pair)의 간극을 최대화하도록 모델을 학습시키는 대조 학습(Contrastive Learning) 기법으로 나뉜다.

[0004] 최근 컴퓨터 비전 분야에도 트랜스포머 기반의 모델에 대한 연구가 활발하게 진행되고 있으며, 그 중 하나가 비전 트랜스포머(Vision Transformer)이다.

[0005] 트랜스포머 모델은 기존의 순환 신경망 기반 모델의 문제점들을 해결한 모델로서, 자연어 처리분야에서 가장 좋은 성능을 보이고 있다. 도 1은 트랜스포머 모델을 도시한 구조도이다. 특히, 트랜스포머 모델은 순환 신경망에 어텐션 알고리즘을 사용하여 구성된 모델이다. 도 2는 트랜스포머 모델의 멀티 헤드 어텐션 구조를 도시한 것이다. 도 2에 도시된 바와 같이, 트랜스포머 모델이 다양한 방법으로 어텐션을 시도하도록 유도하기 위하여, h번 셀프 어텐션 알고리즘을 수행한 뒤 그 정보들을 취합하여 최종 표현을 찾아주는 멀티헤드 셀프 어텐션 기법이 제안되기도 하였다.

[0006] 도 3은 비전 트랜스포머 모델을 도시한 구조도이다. 비전 트랜스포머 모델은 기존 트랜스포머 모델의 구조를 그대로 사용하여 이미지 분류 문제를 해결하기 위하여 사용되는데, 도 3에 도시된 바와 같이 이미지를 패치(Patch)로 잘라 패치들을 하나의 시계열 데이터로 본다. 패치들을 트랜스포머 모델에 입력으로 줄 수 있는 차원으로 투사하여 주는 패치 임베딩(Patch Embedding)을 진행한 다음, 이미지 전체에 대한 내용을 취합하는 클래스 토큰(Class Token)을 추가하여 준다. 그 다음, 기존의 트랜스포머 모델이 입력 데이터에 시간성을 부여하는 것과 유사하게 패치들의 위치를 모델이 알 수 있도록 위치 임베딩을 진행한다. 그렇게 만들어진 입력 데이터를 트랜스포머에 넣어서 나오는 출력 벡터들 중 클래스 토큰에 대응하는 벡터 하나만 사용하여 최종 이미지에 대한 시각적 표현 벡터를 얻게 된다.

[0007] 전술한 방법은 이미지의 픽셀들을 하나의 시계열 데이터로 보고 트랜스포머 모델을 적용시키지 못했던 이유인 트랜스포머의 계산량과 메모리 사용량이 시계열 데이터 길이의 비례하는 점 때문에 효율적이지 못하다는 문제를 해결할 수 있게 된다. 하지만, 트랜스포머의 계산량과 메모리 사용량을 고려하여 패치 크기를 조절해야 하는 문제와 학습 데이터가 커야지만 합성곱 신경망 기반 모델보다 좋을 성능을 보인다는 문제가 있다.

[0008] 한편, 스윈 트랜스포머(Swin Transformer)는 비전 트랜스포머에 합성곱 신경망(CNN) 모델이 갖고 있는 장점인 계층적 표현을 적용시킨 모델이다. 도 4는 스윈 트랜스포머 모델을 도시한 구조도이다. 도 4에 도시된 바와 같이, 스윈 트랜스포머는 로컬 어텐션(Local Attention) 기법을 사용하여 트랜스포머가 구해야 하는 어텐션 맵(Attention Map)의 크기를 패치 수에 비례하지 않고 패치 묶음의 단위인 윈도우(Window)의 크기에 비례하도록 해주는 윈도우 멀티헤드 셀프 어텐션 방법을 제공한다. 윈도우 멀티헤드 셀프 어텐션 방법에서는 모든 패치간의 상관 관계를 고려하지 않고 윈도우 안에 있는 패치간의 상관 관계만 고려한다. 이렇게 하게 되면, 서로 다른 윈도우에 있는 패치간의 상관 관계는 고려하지 않게 되므로, 이를 해결하기 위하여 이동 윈도우(Shifted-Window) 기법을 제안한다.

[0009] 도 5는 이동 윈도우 기법을 설명하기 위하여 도시한 모식도이다. 도 5에 도시된 바와 같이, 이동 윈도우 기법은 매층마다 윈도우의 위치를 변환하여 주는 방법이다.

[0010] 스윈 트랜스포머 모델은 패치 머징(Patch Merging) 기법을 사용하여 단계별로 윈도우가 처리하는 이미지의 범위를 늘린다. 패치 머징은 인접한 패치 4개를 하나의 패치로 합쳐주는 방법이다. 이러한 방법을 적용시킴으로써, 스윈 트랜스포머 모델은 기존의 비전 트랜스포머보다 작은 패치로 이미지를 자름으로써 여러 문제에서 더 좋은

성능을 내게 된다.

[0011] 한편, 비전 트랜스포머를 학습하기 위해서는 많은 양의 데이터가 필요하지만, 실제로 많은 학습 데이터를 갖기는 어렵다. 이에 자기지도학습을 통해 학습 데이터의 양에 대한 문제를 해결하고자 하는 연구가 많이 진행되었다. 자기지도학습 비전 트랜스포머(SiT)는 이미지의 회전률을 예측하고 Contrastive Loss를 최소화하고 마지막으로 손상된 이미지를 복원시킨다. 도 6은 자기지도학습 비전 트랜스포머 모델을 도시한 구조도이다. 도 6을 참조하면, 자기지도학습 비전 트랜스포머는 회전률 토큰(Rotation Token)과 CL(Contrastive Loss) 토큰을 입력 이미지에 추가하여 준다. 또한 기존에는 클래스 토큰에서 나오는 출력만 사용하여 이미지를 분류해서 패치별로 나오는 출력은 사용하지 않았다면, SiT는 패치별로 나오는 출력은 해당 패치를 다시 복구하는데 사용된다. SiT 모델이 제안하는 자기지도학습 방법은 다음과 같다. 먼저, 이미지를 증강하고, 증강된 이미지를 0도, 90도, 180도, 270도 중 랜덤하게 하나로 회전한 뒤 패치들에 임의로 노이즈를 섞는다. 이렇게 증강된 이미지를 SiT 모델에 넣어서 나온 출력값 중 회전률 토큰으로는 랜덤하게 회전한 각도를 예측하고, CL 토큰으로는 손실함수를 최소화하고 패치별로 나온 출력은 본래 이미지의 패치들을 복원하도록 모델을 학습한다. 위와 같은 방식을 통해 학습된 비전 트랜스포머 모델을 선형 검증(Linear Evaluation) 기법을 통해 확인했을 때, 기존의 CL 기법을 통해 학습된 모델보다 좋은 성능을 보인다.

[0012] 하지만, 비전 트랜스포머는 설정한 패치보다 작은 크기의 패치 간의 상관 관계가 고려되기 어려우며, 학습하기 위하여 많은 데이터를 필요로 하는 문제점이 있다.

선행기술문헌

특허문헌

[0013] (특허문헌 0001) 한국등록특허공보 제 10-2189373호

(특허문헌 0002) 한국등록특허공보 제 10-2306344호

(특허문헌 0003) 한국공개특허공보 제 10-2021-0043995호

(특허문헌 0004) 한국공개특허공보 제 10-2021-0152687호

발명의 내용

해결하려는 과제

[0014] 전술한 문제점을 해결하기 위한 본 발명은 로컬 어텐션 알고리즘을 적용한 스윈 트랜스포머(Swin Transformer) 모델에 자기지도학습 알고리즘을 적용하여 트랜스포머의 성능을 향상시키는 방법을 제공하는 것을 목적으로 한다.

[0015] 또한, 본 발명은 윈도우 토큰(Window Token)을 적용한 다중 목적 함수를 활용하여 스윈 트랜스포머 모델을 학습함으로써 비전 트랜스포머의 성능을 향상시키는 방법을 제공하는 것을 다른 목적으로 한다.

과제의 해결 수단

[0016] 전술한 기술적 과제를 달성하기 위한 본 발명의 제1 특징에 따른 자기지도 스윈 트랜스포머 모델 구조는, 컴퓨터 시스템의 프로그램에 의해 구현되는 구조로서, 이미지를 입력받는 이미지 입력 모듈; 입력된 이미지를 사전 설정된 크기로 분할하여 패치들을 생성하고, 상기 패치들을 사전 설정된 크기로 묶어 윈도우들을 생성하는 패치 생성 모듈; 각 패치들을 임베딩하여 벡터로 변환하는 패치 임베딩 모듈; 윈도우의 위치를 변환시키는 이동 윈도우 멀티헤드 셀프 어텐션(SW-MSA) 모듈; 로컬 어텐션 알고리즘을 사용하여 어텐션 맵의 크기를 윈도우의 크기에 비례하도록 해주는 윈도우 멀티헤드 셀프 어텐션(W-MAS) 모듈; 인접한 패치들을 하나의 패치로 합쳐주는 패치 머징 모듈; 을 구비하고, 상기 SW-MSA 모듈 및 W-MAS 모듈은 쌍으로 이루어져 각 단계별로 반복적으로 실행되고, 마지막 단계에는 하나의 윈도우가 전체 이미지를 감싸는 것을 특징으로 한다.

[0017] 전술한 제1 특징에 따른 자기지도 스윈 트랜스포머 모델 구조에 있어서, 상기 SW-MSA 모듈은 윈도우에서 윈도우 토큰을 제외한 패치들을 사용하여 셀프 어텐션 알고리즘을 수행하여 윈도우 간의 정보를 공유하도록 구성된 것이 바람직하다.

[0018] 전술한 제1 특징에 따른 자기지도 스윈 트랜스포머 모델 구조에 있어서, 상기 W-MAS 모듈은, 윈도우 토큰을 패치들과 같이 사용하여 셀프 어텐션 알고리즘을 수행하여 윈도우 토큰에 윈도우의 정보를 집어넣도록 구성된 것이 바람직하다.

[0019] 전술한 제1 특징에 따른 자기지도 스윈 트랜스포머 모델 구조에 있어서, 상기 패치 머징 모듈은, SW-MSA 모듈과 W-MSA 모듈의 쌍이 실행된 후 다음 단계로 넘어갈 때 인접한 4개의 패치를 하나로 만드는 패치 임베딩을 하고, 근접한 윈도우 토큰 4개를 하나로 만드는 토큰 임베딩을 하는 것이 바람직하다.

[0020] 전술한 제1 특징에 따른 자기지도 스윈 트랜스포머 모델 구조에 있어서, 상기 패치 생성 모듈은, 각 윈도우에 대해 회전률 토큰과 Contrastive Loss 토큰을 부여하는 것이 바람직하다.

[0021] 본 발명의 제2 특징에 따른 자기지도 스윈 트랜스포머 모델 구조의 학습 방법은, (a) 입력된 이미지에 대하여 사전 설정된 이미지 증가 기법을 사용하여 2개의 이미지로 증가시키고, 증가된 2개의 이미지를 랜덤하게 회전시킨 후, 임의로 망가뜨려, 입력된 이미지를 2개의 이미지로 증가시키는 단계; (b) 2개의 증가 이미지를 자기지도 스윈 트랜스포머 모델에 입력하는 단계; (c) 상기 자기지도 스윈 트랜스포머 모델에서 출력으로 토큰값들을 사용하여 이미지 회전률과 Contrastive Loss를 구하는 단계; 및 (d) 패치값들을 이용하여 본래의 패치로 복원하는 단계;를 구비한다.

발명의 효과

[0022] 본 발명은 윈도우 토큰을 적용한 스윈 트랜스포머와 다중 목적 함수 기반 자기지도 학습 알고리즘을 적용함으로써, 기존의 비전 트랜스포머가 가진 문제점들인 작은 패치간의 상관 관계를 고려하지 못하는 점과 학습하기 위하여 많은 데이터가 필요한 점들을 해결하였다.

[0023] 또한, 전술한 구조를 갖는 본 발명에 따른 자기지도 스윈 트랜스포머 모델 구조는 기존의 모델 구조들에 비하여 회전률 예측 정확도가 높고, 이로 인해 이미지에 대한 시각적 표현을 잘 찾아낼 수 있게 된다. 또한, 본 발명에 따른 자기지도 스윈 트랜스포머 모델 구조는 기존의 모델 구조들에 비하여 선형 검증 실험에서도 우수한 정확도를 가질 뿐만 아니라, 사후 학습 실험에서도 우수한 성능을 보여주었다.

도면의 간단한 설명

[0024] 도 1은 트랜스포머 모델을 도시한 구조도이다.

도 2는 트랜스포머 모델의 멀티 헤드 어텐션 구조를 도시한 것이다.

도 3은 비전 트랜스포머 모델을 도시한 구조도이다.

도 4는 스윈 트랜스포머 모델을 도시한 구조도이다.

도 5는 이동 윈도우 기법을 설명하기 위하여 도시한 모식도이다.

도 6은 자기지도학습 비전 트랜스포머 모델을 도시한 구조도이다.

도 7은 본 발명에 따른 스윈 트랜스포머 모델 구조에 있어서, 윈도우 토큰 작동 과정을 설명하기 위하여 도시한 이미지들이다.

도 8은 본 발명의 바람직한 실시예에 따른 자기지도 스윈 트랜스포머 모델을 도시한 구조도이다.

도 9는 본 발명의 바람직한 실시예에 따른 자기지도 스윈 트랜스포머 모델에 대한 학습 방법을 도시한 순차적으로 도시한 순서도이며, 도 10은 도 9의 방법을 구현한 알고리즘이다.

발명을 실시하기 위한 구체적인 내용

[0025] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 윈도우 토큰을 적용한 스윈 트랜스포머 모델 구조 및 상기 스윈 트랜스포머 모델의 학습 방법에 대하여 구체적으로 설명한다.

[0026] 기존의 스윈 트랜스포머는 패치들을 작게 잘라 윈도우별로 셀프 어텐션 알고리즘을 수행하며 지속적으로 더 나은 패치에 대한 표현을 찾는다. 그 중간에 패치 머징을 넣어 줌으로써 윈도우가 처리하는 범위를 지속적으로 늘리며, 마지막 단계에서 하나의 윈도우가 이미지 전체를 처리할 수 있게 만든다. 마지막으로 모델이 찾는 이미지에 대한 최종 표현은 각 패치별로 나오는 표현을 평균내서 사용한다. 전술한 기존의 스윈 트랜스포머 모델은 다중 목적 함수를 활용하여 학습시킬 수 없는 문제점이 있다. 따라서, 본 발명에 따른 스윈 트랜스포머 모델 구조

는 윈도우 토큰 기법을 적용한 것을 특징으로 한다.

[0027] 클래스 토큰은 구글에서 제안한 BERT 모델에서 사용된 뒤 트랜스포머 모델을 분류모델로 사용할 때 자주 사용된다. 클래스 토큰에 시계열 데이터가 가지고 있는 정보들 중 분류 문제에 필요한 정보를 갖는 층을 지나면서 지속적으로 축적하여, 최종적으로 분류를 위한 선형층(Linear Layer)에는 클래스 토큰에서 나온 출력만 사용한다. 윈도우 토큰은 로컬 어텐션 알고리즘을 적용한 스윈 트랜스포머 모델에 맞춰 설계된 클래스 토큰으로 작동한다.

[0028] 도 7은 본 발명에 따른 스윈 트랜스포머 모델 구조에 있어서, 윈도우 토큰 작동 과정을 설명하기 위하여 도시한 이미지들이다. 도 7을 참조하면, 사진을 패치로 자른 후 패치의 수를 N*N이라 하고, 윈도우 크기를 M*M이라고 할 때, 총 윈도우 개수는 N/M * N/M 이 되며, 윈도우 별로 윈도우 토큰을 부여한다. 해당 윈도우 토큰은 특정 윈도우에 대한 정보 중 특정 문제에 필요한 정보만 축약하는 토큰이다. 하지만, 스윈 트랜스포머 모델에서는 단계별로 패치 머징이 일어나서 윈도우가 처리하는 범위는 지속적으로 커지게 되는 문제가 발생된다. 이때 도 7과 같이 윈도우 머징 기법을 같이 사용하여 해당 문제를 해결하게 된다. 윈도우 머징 기법은 수학식 1을 통해서 이루어진다. 여기서, ωn은 새롭게 만들어진 윈도우 토큰이며, ω1, ω2, ω3, ω4는 그 전 단계의 인접한 윈도우 토큰들이다.

수학식 1

[0029]
$$w_n = FC(concat(w_1, w_2, w_3, w_4))$$

$$w_1, w_2, w_3, w_4 \in R^N; w_n \in R^{2N}$$

[0030] 이하, 본 발명에 따른 자기지도학습 방법을 통해 학습하는 스윈 트랜스포머 모델 구조를 설명한다.

[0031] 도 8은 본 발명의 바람직한 실시예에 따른 자기지도 스윈 트랜스포머 모델을 도시한 구조도이다. 도 8을 참조하면, 본 발명에 따른 자기지도 스윈 트랜스포머 모델은, 컴퓨터 시스템에서 실행되는 프로그램 또는 소프트웨어에 의해 구현된 모델로서, 이미지 입력 모듈, 패치 생성 모듈, 패치 임베딩 모듈, 이동 윈도우 멀티헤드 셀프 어텐션(SW-MSA) 모듈, 윈도우 멀티헤드 셀프 어텐션(W-MSA) 모듈 및 패치 머징 모듈을 구비하고, 상기 SW-MSA 모듈 및 W-MAS 모듈은 쌍으로 이루어져 각 단계별로 반복적으로 실행되고, 마지막 단계에는 하나의 윈도우가 전체 이미지를 감싸는 것을 특징으로 한다.

[0032] 상기 이미지 입력 모듈은, 224*224*3의 이미지를 입력받는다.

[0033] 상기 패치 생성 모듈은 이미지 입력 모듈을 통해 입력된 이미지를 4*4*3 패치로 잘라 56*56 개의 패치를 생성하고, 해당 패치들을 7*7 크기를 갖는 윈도우로 묶는다. 그러면, 총 8*8 개의 윈도우가 생성되며 윈도우마다 개별적으로 회전률 토큰과 CL 토큰을 부여한다. 상기 패치들은 패치 임베딩(Patch Embedding) 모듈을 통해 96 차원을 가진 벡터로 변환된다.

[0034] 그후 사용되는 층의 구조는 W-MSA 모듈과 SW-MSA 모듈로 이루어진다. 상기 이동 윈도우 멀티헤드 셀프 어텐션(SW-MSA) 모듈은 윈도우 멀티헤드 셀프 어텐션(W-MSA) 모듈보다 먼저 실행됨으로써, SW-MSA 모듈은 움직여진 윈도우에서 윈도우 토큰을 제외한 패치들을 사용하여 셀프 어텐션 알고리즘을 수행하고, 이는 윈도우 간의 정보를 공유하게 한다.

[0035] 다음, W-MSA 모듈을 수행할 때 윈도우 토큰을 패치들과 같이 사용하여 셀프 어텐션 모듈을 진행함으로써 윈도우 토큰에 윈도우의 정보를 집어넣어 준다. 본 발명의 따른 바람직한 실시예에 따른 자기지도 스윈 트랜스포머 모델은 총 12개의 층으로 이루어지며 4개의 단계로 나뉘어진다. 매 단계에서 다음 단계로 넘어갈 때, 인접한 4개의 패치를 하나로 만드는 패치 임베딩을 거치고, 이때 근접한 윈도우 토큰 4개를 하나로 만드는 토큰 임베딩도 같이 진행한다. 각 단계에는 단계별로 SW-MSA, W-MSA를 하나의 쌍으로 하여, 2개, 2개, 6개, 2개의 층을 가지고 있으며 마지막 단계에는 하나의 윈도우가 전체 이미지를 다 감싸게 된다.

[0036] 이하, 본 발명에 따른 자기지도 스윈 트랜스포머 모델에 대한 학습 방법에 대하여 설명한다. 도 9는 본 발명의 바람직한 실시예에 따른 자기지도 스윈 트랜스포머 모델에 대한 학습 방법을 도시한 순차적으로 도시한 순서도이며, 도 10은 도 9의 방법을 구현한 알고리즘이다.

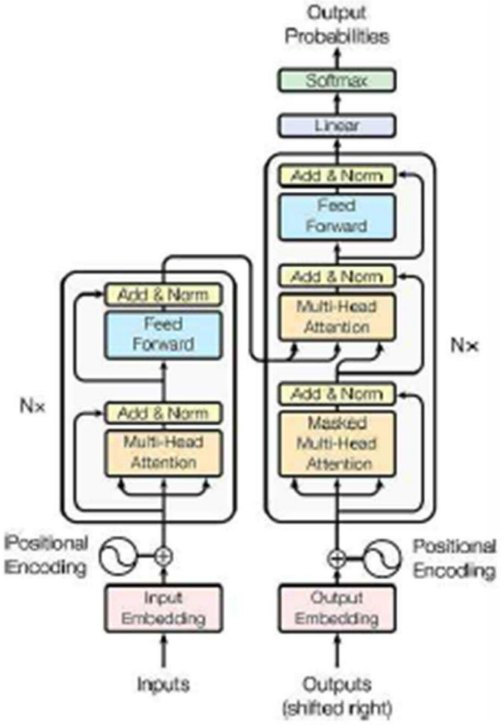
[0037] 도 9 및 도 10을 참조하면, 본 발명에 따른 자기지도 스윈 트랜스포머 모델에 대한 학습 방법은, 먼저 이미지가

입력으로 들어오면 사전 설정된 이미지 증가 기법을 사용하여 이미지를 2가지 버전으로 증가시킨다. 증가된 2개의 이미지에 랜덤한 회전을 가해준다. 다음, 2개의 이미지를 임의로 망가뜨린 후 전술한 본 발명에 따른 자기지도 스윈 트랜스포머 모델에 넣어준다. 그 후, 본 발명에 따른 스윈 트랜스포머 모델에서 출력으로 주는 토큰 값들을 사용하여 이미지 회전률과 Contrastive Loss를 구해주고 패치 값들을 사용하여 본래의 패치로 복원한다.

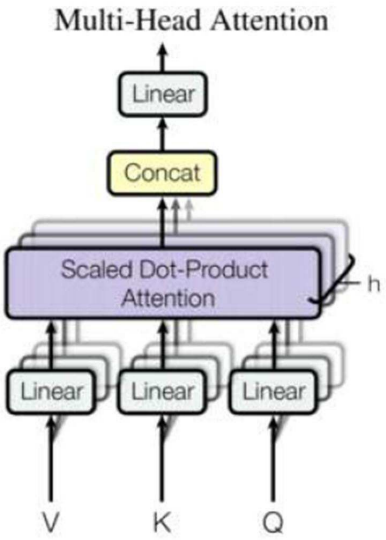
[0038] 이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

도면

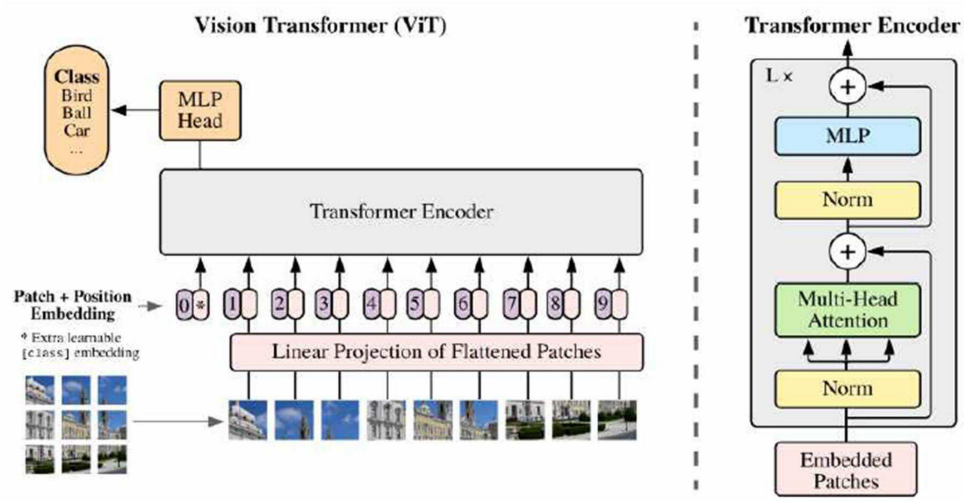
도면1



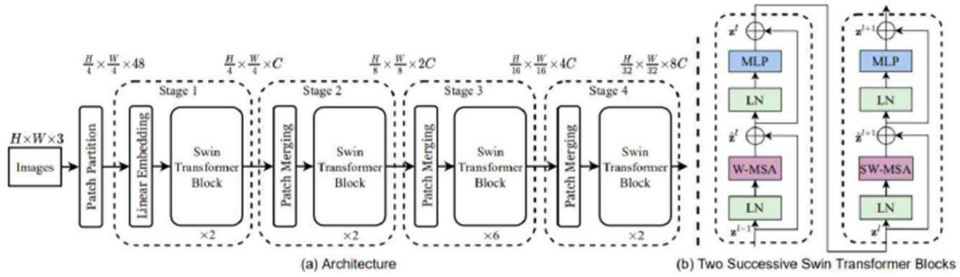
도면2


Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

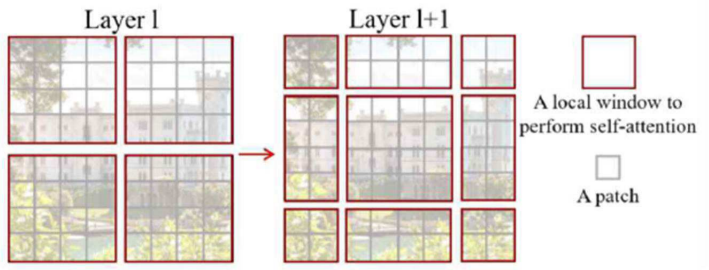
도면3


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

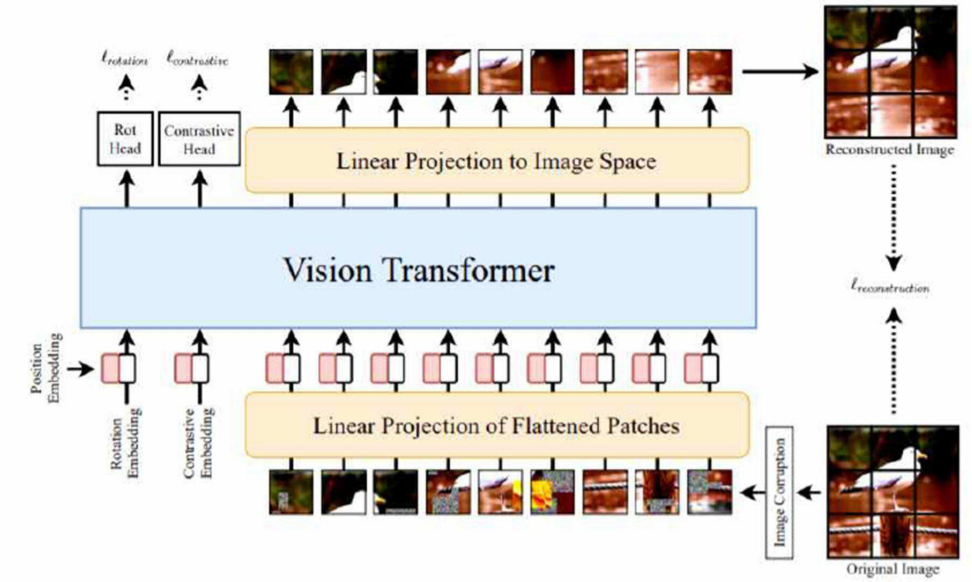
도면4


H×W×3
Images
Patch Partition
Stage 1
Linear Embedding
Swin Transformer Block
×2
Stage 2
Patch Merging
Swin Transformer Block
×2
Stage 3
Patch Merging
Swin Transformer Block
×6
Stage 4
Patch Merging
Swin Transformer Block
×2
(a) Architecture
MLP
LN
W-MSA
LN
MLP
LN
SW-MSA
LN
(b) Two Successive Swin Transformer Blocks

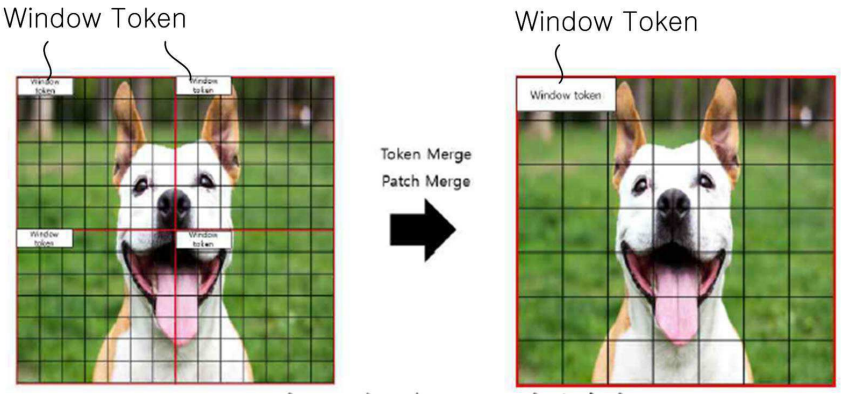
도면5


Layer l
Layer l+1
A local window to perform self-attention
A patch

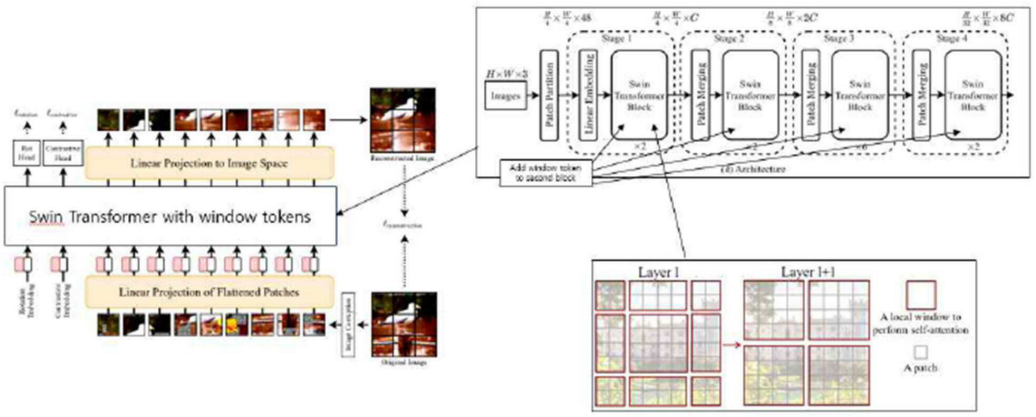
도면6


$\ell_{rotation}$
$\ell_{contrastive}$
Rot Head
Contrastive Head
Linear Projection to Image Space
Reconstructed Image
Vision Transformer
$\ell_{reconstruction}$
Position Embedding
Rotation Embedding
Contrastive Embedding
Linear Projection of Flattened Patches
Image Corruption
Original Image

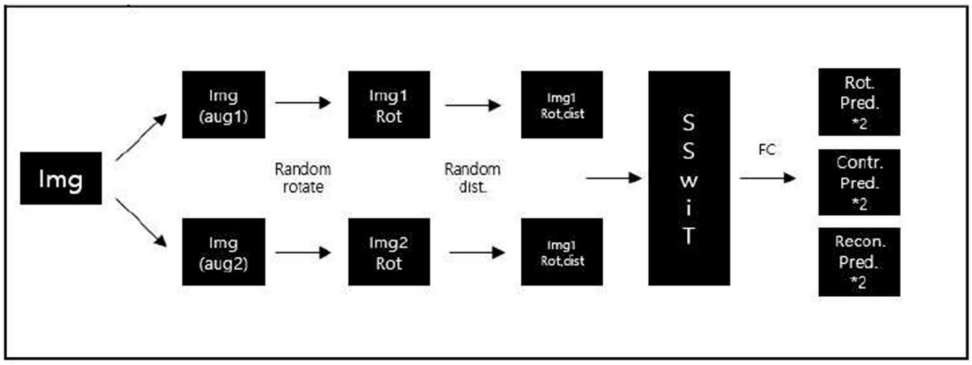
도면7


Window Token
Token Merge
Patch Merge
Window Token
Window token

도면8


Linear Projection to Image Space
Swin Transformer with window tokens
Linear Projection of Flattened Patches
Images
Patch Partition
Linear Embedding
Swin Transformer Block
Patch Merging
Stage 1
Stage 2
Stage 3
Stage 4
Layer l
Layer l+1
A local window to perform self-attention
A patch

도면9


Img
Img (aug1)
Img (aug2)
Random rotate
Img1 Rot
Img2 Rot
Random dist.
Img1 Rot.dist
Img1 Rot.dist
SSwiT
FC
Rot. Pred. *2
Contr. Pred. *2
Recon. Pred. *2

도면10

Algorithm 2. SSwiT 학습과정

input : mini batch N, transformer encoder f, encoder g, h, t, augmentation candidate A, distortion candidate D

for mini-batch x_k, $(k = 1, 2, \ldots N)$ **do**

for k=1 to N **do**

two augmentation $a \rightarrow A$, $a' \rightarrow A$ # a, a' : augmentation

random rotation $r \rightarrow 0, 90, 180, 270$

two distortion $d \rightarrow D, d' \rightarrow D$ # d,d' : distortion

$x_{2k-1} = a(x_k)$ #first augmentation

$r_{2k-1}, rot_{true_{2k-1}} = r(x_{2k-1})$

$d_{2k-1}, recon_{true_{2k-1}} = d(r_{2k-1})$

$x_{2k} = a'(x_k)$ #second augmentation

$r_{2k}, rot_{true_{2k}} = r(x_{2k})$

$d_{2k}, recon_{true_{2k}} = d(r_{2k})$

end for

input = concat(tokens, d)

$repr_{rot}, repr_{cl}, repr_{rec} = f(\in put)$

$rot_{pred} = g(repr_{rot})$

$s = h(repr_{cl})$

$recon_{pred} = t(repr_{rec})$

for i, j = 1 to 2N **do**

$s_{i,j} = z_i^T z_j / (\tau \|z_i\| \|z_j\|)$

end for

Let $l_1(a,b) = -\frac{1}{2N}\sum_{i=1}^{2N}\sum_{j=1} b_{i_j} \log(a_{i_j})$

Let $l_2(i,j) = -\log \frac{\exp(s_{i,j})}{\sum_{k=1[k \neq i]}^{2N} \exp(s_{i,k})}$

Let $l_3(a,b) = \frac{1}{2N}\sum_{i=1}^{2N} |b_i - a_i|$

$loss = \frac{1}{3}(l_1(rot_{pred}, rot_{true}) + \frac{1}{2N}\sum_{k=1}^{N}[l_2(2k-1, 2k) +$

$+ l_2(2k, 2k-1)] + l_3(recon_{pred}, recon_{true}))$

update f & g & h & t networks to minimize loss

end for

return encoder f