



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0083996
(43) 공개일자 2025년06월10일

(51) 국제특허분류(Int. Cl.)
G06V 10/26 (2022.01) G06N 3/08 (2023.01)
G06T 7/11 (2017.01) G06V 10/44 (2022.01)
G06V 10/74 (2022.01) G06V 10/774 (2022.01)
(52) CPC특허분류
G06V 10/26 (2023.08)
G06N 3/08 (2023.01)
(21) 출원번호 10-2024-0065166
(22) 출원일자 2024년05월20일
심사청구일자 2024년05월20일
(30) 우선권주장
1020230172539 2023년12월01일 대한민국(KR)

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
심재형
서울특별시 마포구 만리재로 74, 207동 801호(신공덕동, 신공덕2차삼성래미안)
권민서
서울특별시 영등포구 국제금융로7길 20, 3동 207호(여의도동, 대교아파트)
(74) 대리인
김연권

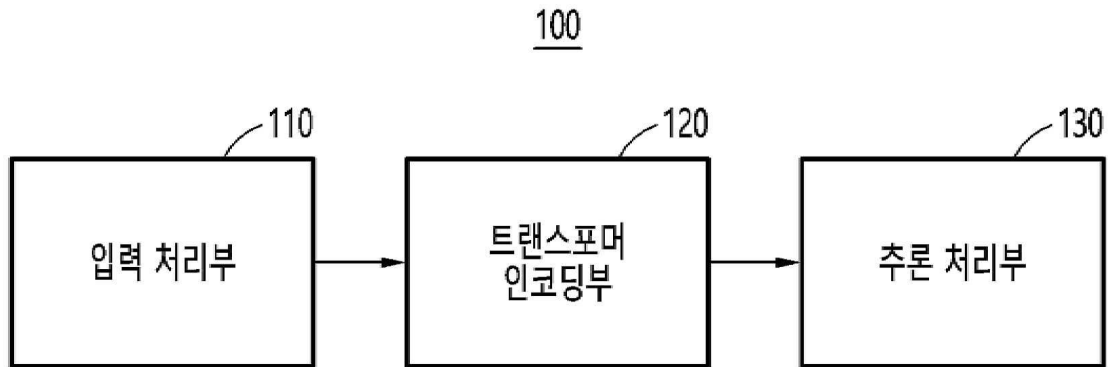
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 토큰 병합을 이용한 비전 트랜스포머 장치 및 방법

(57) 요약

본 발명은 토큰 병합을 이용한 비전 트랜스포머 장치 및 방법에 관한 것으로, 본 발명의 일실시예에 따른 비전 트랜스포머 장치는 인식 대상 이미지를 복수의 패치로 분할하고, 상기 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리하는 입력 처리부, 상기 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 트랜스포머 인코딩부 및 상기 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 상기 인식 대상 이미지에서 객체 인식 추론을 수행하는 추론 처리부를 포함할 수 있다.

대표도 - 도1



(52) CPC특허분류

G06T 7/11 (2017.01)
G06V 10/454 (2023.08)
G06V 10/761 (2023.08)
G06V 10/774 (2023.08)

김효진

대구광역시 수성구 들안로73길 50-7, 101동 803호
 (수성동4가, 수성대림 e-편한세상)

(72) 발명자

권수영

경기도 성남시 분당구 서판교로 29, 921동 104호(
 판교동, 판교원마을한림폴에버아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711179344
과제번호	00155966
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신·방송 연구개발사업-인력양성사업-인공지능융합혁신인재양성
연구과제명	인공지능융합혁신인재양성(이화여자대학교)
과제수행기관명	이화여자대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193622
과제번호	2021-0-02068-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능 혁신 허브 연구 개발
과제수행기관명	이화여자대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

인식 대상 이미지를 복수의 패치로 분할하고, 상기 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리하는 입력 처리부;

상기 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 트랜스포머 인코딩부; 및

상기 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 상기 인식 대상 이미지에서 객체 인식 추론을 수행하는 추론 처리부를 포함하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 2

제1항에 있어서,

상기 트랜스포머 인코딩 처리부는 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 유사 토큰을 선정하고, 상기 선정된 유사 토큰에 대한 정보를 저장하며, 상기 선정된 적어도 하나의 유사 토큰과 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 추가 유사 토큰을 선정하며, 상기 적어도 하나의 추가 유사 토큰에 대한 정보를 저장하고, 상기 선정된 기준 토큰, 상기 선정된 적어도 하나의 유사 토큰 및 상기 선정된 적어도 하나의 추가 유사 토큰을 한번에 병합하여 상기 복수의 트랜스포머 블록을 통해 상기 토큰 병합 처리 결과를 출력하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 3

제2항에 있어서,

상기 트랜스포머 인코딩 처리부는 상기 토큰 매트릭스를 구성하는 토큰들에 대하여 어텐션 분포(attention distribution)에 대한 연산을 수행하고, 상기 연산된 어텐션 분포와 상기 토큰들의 공간 지역성(spatial locality)에 기반하여 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도를 비교하고, 상기 비교된 유사도가 상기 기준 유사도보다 큰 토큰을 상기 적어도 하나의 유사 토큰으로 선정하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 4

제3항에 있어서,

상기 트랜스포머 인코딩 처리부는 상기 토큰 매트릭스를 구성하는 토큰들에 대한 어텐션 점수(score)에 소프트맥스(softmax) 함수를 적용하여 상기 어텐션 분포를 연산하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 5

제1항에 있어서,

상기 트랜스포머 인코딩 처리부는 상기 토큰 매트릭스를 구성하는 토큰들에 대한 토큰 병합 비율을 증가시키도

록 상기 기준 유사도를 감소 설정하고, 상기 토큰 병합 비율을 감소시키도록 상기 기준 유사도를 증가 설정하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 6

제5항에 있어서,

상기 트랜스포머 인코딩 처리부는 상기 인식 대상 이미지에서 배경(background) 이미지 비율을 고려하여 상기 기준 유사도를 감소 설정하고, 상기 인식 대상 이미지에서 다수의 식별 대상 객체의 수를 고려하여 상기 기준 유사도를 증가 설정하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 7

제1항에 있어서,

상기 추론 처리부는 상기 토큰 병합 처리 결과에서 균일한 색상을 가진 인접한 배경이 하나의 토큰으로 병합됨에 따라 상기 토큰 매트릭스를 구성하는 토큰의 수가 감소하고, 상기 감소된 토큰에 기반하여 다층 퍼셉트론(multilayer perceptron)의 헤드 처리 통해 상기 인식 대상 이미지에서 객체에 대한 클래스를 분류하는 상기 객체 인식 추론을 수행하는 것을 특징으로 하는

비전 트랜스포머 장치.

청구항 8

입력 처리부에서, 인식 대상 이미지를 복수의 패치로 분할하고, 상기 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리하는 단계;

트랜스포머 인코딩부에서, 상기 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 단계; 및

추론 처리부에서, 상기 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 상기 인식 대상 이미지에서 객체 인식 추론을 수행하는 단계를 포함하는 것을 특징으로 하는

비전 트랜스포머 방법.

청구항 9

제8항에 있어서,

상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 단계는,

상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 유사 토큰을 선정하고, 상기 선정된 유사 토큰에 대한 정보를 저장하며, 상기 선정된 적어도 하나의 유사 토큰과 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 추가 유사 토큰을 선정하며, 상기 적어도 하나의 추가 유사 토큰에 대한 정보를 저장하고, 상기 선정된 기준 토큰, 상기 선정된 적어도 하나의 유사 토큰 및 상기 선정된 적어도 하나의 추가 유사 토큰을 한번에 병합하여 상기 복수의 트랜스포머 블록을 통해 상기 토큰 병합 처리 결과를 출력하는 단계를 포함하는 것을 특징으로 하는

비전 트랜스포머 방법.

청구항 10

제9항에 있어서,

상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 단계는,

상기 토큰 매트릭스를 구성하는 토큰들에 대하여 어텐션 분포(attention distribution)에 대한 연산을 수행하고, 상기 연산된 어텐션 분포와 상기 토큰들의 공간 지역성(spatial locality)에 기반하여 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도를 비교하고, 상기 비교된 유사도가 상기 기준 유사도보다 큰 토큰을 상기 적어도 하나의 유사 토큰으로 선정하는 단계를 포함하는 것을 특징으로 하는

비전 트랜스포머 방법.

발명의 설명

기술 분야

[0001] 본 발명은 토큰 병합을 이용한 비전 트랜스포머 장치 및 방법에 관한 것으로, 보다 구체적으로, 이미지 데이터의 특징을 활용해 인접한 토큰들끼리 유사도를 비교하여 추가적인 모델 학습 없이도 한 번의 토큰 병합(token merging at once)을 수행하여 비전 트랜스포머에 기반한 이미지 인식 및 처리와 관련된 추론 가속화를 구현하는 기술에 관한 것이다.

배경 기술

[0002] 문장의 구조와 의미적 관계를 파악하고 생성하는데, 강력한 성능을 보이는 트랜스포머(transformer) 모델이 등장하면서, 자연어 처리 분야 뿐만 아니라 컴퓨터 비전, 음성 인식 분야 등 다양한 분야에서 트랜스포머 기반의 모델들이 괄목할 만한 성능을 보이고 있다.

[0003] 특히, 그중 비전 트랜스포머(vision transformer, ViT) 모델은 이미지 분류 작업에서 고전적인 CNN 모델들과 비교하여 우수한 성능을 보여준다.

[0004] 하지만, 비전 트랜스포머 모델의 뛰어난 정확도에도 불구하고, 비전 트랜스포머는 거대한 모델 크기로 인한 추론 시간의 지연이라는 문제점이 존재한다.

[0005] 이를 해결하기 위해서 이미지 내의 불필요한 토큰들을 가지치기해서 처리하면, 처리해야할 연산량을 줄이는 토큰 가지치기(token pruning) 기법을 비전 트랜스포머에 적용하는 연구들이 제안되어 왔다.

[0006] 다만, 종래 기술에서는 토큰을 제거함에 따라 정보 손실이 일어나기 때문에 정확도가 떨어지고, 모델의 성능을 높이기 위해 미세조정(fine-tuning)이 요구되었다.

[0007] 토큰 가지치기 기법의 문제점을 보완하기 위해서 토큰 병합(token merging)이 제안되었다.

[0008] 종래 기술에서는 토큰 병합에서 유사한 토큰들을 점진적으로 결합하여 총 토큰의 수를 줄이고 정보 손실을 방지한다.

[0009] 다만, 복수의 트랜스포머 블록(transformer block)에 걸친 점진적 병합이 요구됨에 따라 추론 속도가 가속화되는데 한계점이 존재한다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 한국공개특허 제10-2023-0149554호, "비전 트랜스포머의 정규화 장치 및 방법"
(특허문헌 0002) 한국공개특허 제10-2023-0085495호, "이미지 적응적인 힌트 생성을 사용하는 이미지 아웃페인팅 방법 및 그 장치"
(특허문헌 0003) 한국등록특허 제10-2479817호, "소규모 데이터셋을 위한 비전 트랜스포머 장치 및 그 동작 방법"

발명의 내용

해결하려는 과제

[0011] 본 발명은 이미지 데이터의 특징을 활용해 인접한 토큰들끼리 유사도를 비교하여 추가적인 모델 학습 없이도 한번의 토큰 병합(token merging at once)을 수행하여 비전 트랜스포머에 기반한 이미지 인식 및 처리와 관련된 추론 가속화를 구현하는 비전 트랜스포머 장치 및 방법을 제공하는 것을 목적으로 한다.

[0012] 본 발명은 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서 토큰 매트릭스(matrix)를 구성하는 복수의 토큰 중 어느 하나의 토큰을 기준으로 인접한 토큰들과의 유사도를 비교하여 연쇄적으로 토큰 병합을 진행함에 높은 정확도를 유지하면서 전반적인 추론 속도를 개선하는 비전 트랜스포머 장치 및 방법을 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0013] 본 발명의 일실시예에 따른 비전 트랜스포머 장치는 인식 대상 이미지를 복수의 패치로 분할하고, 상기 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리하는 입력 처리부, 상기 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 트랜스포머 인코딩부 및 상기 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 상기 인식 대상 이미지에서 객체 인식 추론을 수행하는 추론 처리부를 포함할 수 있다.

[0014] 상기 트랜스포머 인코딩 처리부는 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 유사 토큰을 선정하고, 상기 선정된 유사 토큰에 대한 정보를 저장하며, 상기 선정된 적어도 하나의 유사 토큰과 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 추가 유사 토큰을 선정하며, 상기 적어도 하나의 추가 유사 토큰에 대한 정보를 저장하고, 상기 선정된 기준 토큰, 상기 선정된 적어도 하나의 유사 토큰 및 상기 선정된 적어도 하나의 추가 유사 토큰을 한번에 병합하여 상기 복수의 트랜스포머 블록을 통해 상기 토큰 병합 처리 결과를 출력할 수 있다.

[0015] 상기 트랜스포머 인코딩 처리부는 상기 토큰 매트릭스를 구성하는 토큰들에 대하여 어텐션 분포(attention distribution)에 대한 연산을 수행하고, 상기 연산된 어텐션 분포와 상기 토큰들의 공간 지역성(spatial locality)에 기반하여 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도를 비교하고, 상기 비교된 유사도가 상기 기준 유사도보다 큰 토큰을 상기 적어도 하나의 유사 토큰으로 선정할 수 있다.

[0016] 상기 트랜스포머 인코딩 처리부는 상기 토큰 매트릭스를 구성하는 토큰들에 대한 어텐션 점수(score)에 소프트맥스(softmax) 함수를 적용하여 상기 어텐션 분포를 연산할 수 있다.

[0017] 상기 트랜스포머 인코딩 처리부는 상기 토큰 매트릭스를 구성하는 토큰들에 대한 토큰 병합 비율을 증가시키도록 상기 기준 유사도를 감소 설정하고, 상기 토큰 병합 비율을 감소시키도록 상기 기준 유사도를 증가 설정할 수 있다.

[0018] 상기 트랜스포머 인코딩 처리부는 상기 인식 대상 이미지에서 배경(background) 이미지 비율을 고려하여 상기 기준 유사도를 감소 설정하고, 상기 인식 대상 이미지에서 다수의 식별 대상 객체의 수를 고려하여 상기 기준 유사도를 증가 설정할 수 있다.

[0019] 상기 추론 처리부는 상기 토큰 병합 처리 결과에서 균일한 색상을 가진 인접한 배경이 하나의 토큰으로 병합됨에 따라 상기 토큰 매트릭스를 구성하는 토큰의 수가 감소하고, 상기 감소된 토큰에 기반하여 다층 퍼셉트론(multilayer perceptron)의 헤드 처리 통해 상기 인식 대상 이미지에서 객체에 대한 클래스를 분류하는 상기 객체 인식 추론을 수행할 수 있다.

[0020] 본 발명의 일실시예에 따르면 비전 트랜스포머 방법은 입력 처리부에서, 인식 대상 이미지를 복수의 패치로 분할하고, 상기 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리하는 단계, 트랜스포머 인코

딩부에서, 상기 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 단계 및 추론 처리부에서, 상기 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 상기 인식 대상 이미지에서 객체 인식 추론을 수행하는 단계를 포함할 수 있다.

[0021] 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 단계는, 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 유사 토큰을 선정하고, 상기 선정된 유사 토큰에 대한 정보를 저장하며, 상기 선정된 적어도 하나의 유사 토큰과 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 추가 유사 토큰을 선정하며, 상기 적어도 하나의 추가 유사 토큰에 대한 정보를 저장하고, 상기 선정된 기준 토큰, 상기 선정된 적어도 하나의 유사 토큰 및 상기 선정된 적어도 하나의 추가 유사 토큰을 한번에 병합하여 상기 복수의 트랜스포머 블록을 통해 상기 토큰 병합 처리 결과를 출력하는 단계를 포함할 수 있다.

[0022] 상기 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 상기 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행하는 단계는, 상기 토큰 매트릭스를 구성하는 토큰들에 대하여 어텐션 분포(attention distribution)에 대한 연산을 수행하고, 상기 연산된 어텐션 분포와 상기 토큰들의 공간 지역성(spatial locality)에 기반하여 상기 선정된 기준 토큰과 상기 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도를 비교하고, 상기 비교된 유사도가 상기 기준 유사도보다 큰 토큰을 상기 적어도 하나의 유사 토큰으로 선정하는 단계를 포함할 수 있다.

발명의 효과

[0023] 본 발명은 이미지 데이터의 특징을 활용해 인접한 토큰들끼리 유사도를 비교하여 추가적인 모델 학습 없이도 한 번의 토큰 병합(token merging at once)을 수행하여 비전 트랜스포머에 기반한 이미지 인식 및 처리와 관련된 추론 가속화를 구현하는 비전 트랜스포머 장치 및 방법을 제공할 수 있다.

[0024] 본 발명은 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서 토큰 매트릭스(matrix)를 구성하는 복수의 토큰 중 어느 하나의 토큰을 기준으로 인접한 토큰들과의 유사도를 비교하여 연쇄적으로 토큰 병합을 진행함에 높은 정확도를 유지하면서 전반적인 추론 속도를 개선하는 비전 트랜스포머 장치 및 방법을 제공할 수 있다.

도면의 간단한 설명

[0025] 도 1은 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 장치를 설명하는 도면이다.

도 2는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부의 토큰 병합 블록을 설명하는 도면이다.

도 3 및 도 4는 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법을 설명하는 도면이다.

도 5a 내지 도 5e는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서 토큰 병합 절차를 설명하는 도면이다.

도 6은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서 추가 토큰 병합 절차를 설명하는 도면이다.

도 7은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서의 토큰 병합 처리 결과를 설명하는 도면이다.

도 8 및 도 9는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 추론 성능에 대한 시뮬레이션 결과를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 이하, 본 문서의 다양한 실시 예들이 첨부된 도면을 참조하여 기재된다.
- [0027] 실시 예 및 이에 사용된 용어들은 본 문서에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 해당 실시 예의 다양한 변경, 균등물, 및/또는 대체물을 포함하는 것으로 이해되어야 한다.
- [0028] 하기에 다양한 실시 예들을 설명에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.
- [0029] 그리고 후술되는 용어들은 다양한 실시 예들에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [0030] 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.
- [0031] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.
- [0032] 본 문서에서, "A 또는 B" 또는 "A 및/또는 B 중 적어도 하나" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다.
- [0033] "제1," "제2," "첫째," 또는 "둘째," 등의 표현들은 해당 구성요소들을, 순서 또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.
- [0034] 어떤(예: 제1) 구성요소가 다른(예: 제2) 구성요소에 "(기능적으로 또는 통신적으로) 연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 상기 어떤 구성요소가 상기 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제3 구성요소)를 통하여 연결될 수 있다.
- [0035] 본 명세서에서, "~하도록 구성된(또는 설정된)(configured to)"은 상황에 따라, 예를 들면, 하드웨어적 또는 소프트웨어적으로 "~에 적합한," "~하는 능력을 가지는," "~하도록 변경된," "~하도록 만들어진," "~를 할 수 있는," 또는 "~하도록 설계된"과 상호 호환적으로(interchangeably) 사용될 수 있다.
- [0036] 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는" 것을 의미할 수 있다.
- [0037] 예를 들면, 문구 "A, B, 및 C를 수행하도록 구성된(또는 설정된) 프로세서"는 해당 동작을 수행하기 위한 전용 프로세서(예: 임베디드 프로세서), 또는 메모리 장치에 저장된 하나 이상의 소프트웨어 프로그램들을 실행함으로써, 해당 동작들을 수행할 수 있는 범용 프로세서(예: CPU 또는 application processor)를 의미할 수 있다.
- [0038] 또한, '또는' 이라는 용어는 배타적 논리합 'exclusive or' 이기보다는 포함적인 논리합 'inclusive or' 를 의미한다.
- [0039] 즉, 달리 언급되지 않는 한 또는 문맥으로부터 명확하지 않는 한, 'x가 a 또는 b를 이용한다' 라는 표현은 포함적인 자연 순열들(natural inclusive permutations) 중 어느 하나를 의미한다.
- [0040] 이하 사용되는 '...부', '...기' 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어, 또는, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.
- [0042] 도 1은 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 장치를 설명하는 도면이다.
- [0043] 도 1은 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 장치의 구성 요소를 예시한다.
- [0044] 도 1을 참고하면, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 장치(100)는 입력 처리부(110), 트랜스포머 인코딩부(120) 및 추론 처리부(130)를 포함한다.
- [0045] 일례로, 비전 트랜스포머 장치(100)는 입력 이미지를 작은 조각으로 나뉘서 각 조각 내 위치하는 객체를 추론하고, 각 조각에서 추론된 객체에 기반하여 전체 이미지를 인식하는 딥러닝 모델이다.
- [0046] 본 발명의 일실시예에 따른 비전 트랜스포머 장치(100)는 한번에 토큰을 병합하는 프레임 워크에 기반하여 토큰 병합에 소모되는 시간을 감소하여 추가적인 모델 학습 없이도 객체 추론 및 이미지 인식을 수행하는 딥러닝 모델 기반의 장치일 수 있다.
- [0047] 본 발명의 일실시예에 따른 비전 트랜스포머 장치(100)는 딥러닝 모델 추론 전 한번에 토큰 병합을 진행하여 매

트랜스포머 블록마다 병합을 수행하는 종래 기술에 대비하여 토큰 병합에 소모되는 시간을 감소하여 추론 시간을 감소시킬 수 있다.

- [0048] 본 발명의 일실시예에 따른 입력 처리부(110)는 인식 대상 이미지를 복수의 패치로 분할하고, 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리할 수 있다.
- [0049] 일례로, 입력 처리부(110)는 다양한 이미지 인식 및 처리 작업을 수행하는 딥러닝 모델인 비전 트랜스포머 모델에 적용하기 위한 입력 데이터로 이미지를 패치로 나누어서 토큰 형태로 데이터 처리를 수행할 수 있다.
- [0050] 예를 들어, 입력 처리부(110)는 토큰 수 감소 및 정보 손실 방지를 위한 토큰 병합을 위한 입력 데이터 처리를 수행한다.
- [0051] 본 발명의 일실시예에 따르면 트랜스포머 인코딩부(120)는 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행할 수 있다.
- [0052] 트랜스포머 인코딩 처리부(120)는 선정된 기준 토큰과 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 기준 유사도에 기반하여 적어도 하나의 유사 토큰을 선정할 수 있다.
- [0053] 트랜스포머 인코딩 처리부(120)는 선정된 유사 토큰에 대한 정보를 저장하며, 선정된 적어도 하나의 유사 토큰과 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 상기 기준 유사도에 기반하여 적어도 하나의 추가 유사 토큰을 선정할 수 있다.
- [0054] 트랜스포머 인코딩 처리부(120)는 적어도 하나의 추가 유사 토큰에 대한 정보를 저장하고, 선정된 기준 토큰, 선정된 적어도 하나의 유사 토큰 및 선정된 적어도 하나의 추가 유사 토큰을 한번에 병합하여 복수의 트랜스포머 블록을 통해 토큰 병합 처리 결과를 출력할 수 있다.
- [0055] 트랜스포머 인코딩 처리부(120)는 토큰 매트릭스를 구성하는 토큰들에 대하여 어텐션 분포(attention distribution)에 대한 연산을 수행하고, 연산된 어텐션 분포와 토큰들의 공간 지역성(spatial locality)에 기반하여 선정된 기준 토큰과 선정된 기준 토큰에 상, 하, 좌 및 우 중 적어도 하나에 인접한 토큰과의 유사도와 기준 유사도를 비교하고, 비교된 유사도가 기준 유사도보다 큰 토큰을 적어도 하나의 유사 토큰으로 선정할 수 있다.
- [0056] 트랜스포머 인코딩 처리부(120)는 토큰 매트릭스를 구성하는 토큰들에 대한 어텐션 점수(score)에 소프트 맥스(softmax) 함수를 적용하여 어텐션 분포를 연산할 수 있다.
- [0057] 트랜스포머 인코딩 처리부(120)는 토큰 매트릭스를 구성하는 토큰들에 대한 토큰 병합 비율을 증가시키도록 기준 유사도를 감소 설정하고, 토큰 병합 비율을 감소시키도록 기준 유사도를 증가 설정할 수 있다.
- [0058] 트랜스포머 인코딩 처리부(120)는 인식 대상 이미지에서 배경(background) 이미지 비율을 고려하여 기준 유사도를 감소 설정하고, 인식 대상 이미지에서 다수의 식별 대상 객체의 수를 고려하여 기준 유사도를 증가 설정할 수 있다.
- [0059] 트랜스포머 인코딩 처리부(120)는 트랜스포머 블록마다 토큰 병합을 진행하는 종래 기술과 달리, 복수의 트랜스포머 블록 중 어느 하나만 토큰 병합을 진행하여 병합에 필요한 전체적인 소요 시간을 줄일 수 있다.
- [0060] 트랜스포머 인코딩 처리부(120)는 설정에 따른 기준 유사도에 기반하여 토큰 병합되는 정도를 조절할 수 있고, 이 값이 작을수록 더 많은 토큰들이 병합된다.
- [0061] 이미지 상에서 인접한 패치들끼리 비슷한 배경 또는 특징을 가질 확률이 높다는 점이 활용될 수 있다.
- [0062] 균일한 색상을 가진 인접한 배경이 한번에 병합될 수 있고, 이를 바탕으로 토큰끼리의 유사도 비교 횟수를 크게 줄이기 위해 상하좌우로 인접한 토큰들끼리만 비교할 수 있다.
- [0063] 본 발명의 일실시예에 따른 한 번의 토큰 병합은 주어진 이미지에 존재하는 모든 토큰들에 대해 인접한 토큰을 방문한 뒤에 유사한 토큰을 병합하는 과정을 반복적으로 수행한다.
- [0064] 즉, 한 번의 토큰 병합은 주어진 하나의 토큰에 대해 상하좌우로 인접한 4 개의 토큰과의 유사도를 비교하는 과정을 재귀적으로 반복한다.

- [0065] 추론 처리부(130)는 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 인식 대상 이미지에서 객체 인식 추론을 수행할 수 있다.
- [0066] 추론 처리부(130)는 토큰 병합 처리 결과에서 균일한 색상을 가진 인접한 배경이 하나의 토큰으로 병합됨에 따라 토큰 매트릭스를 구성하는 토큰의 수가 감소하고, 감소된 토큰에 기반하여 다층 퍼셉트론(multilayer perceptron)의 헤드 처리 통해 인식 대상 이미지에서 객체에 대한 클래스를 분류하는 객체 인식 추론을 수행할 수 있다.
- [0067] 따라서, 본 발명은 이미지 데이터의 특징을 활용해 인접한 토큰들끼리 유사도를 비교하여 추가적인 모델 학습 없이도 한 번의 토큰 병합(token merging at once)을 수행하여 비전 트랜스포머에 기반한 이미지 인식 및 처리와 관련된 추론 가속화를 구현하는 비전 트랜스포머 장치 및 방법을 제공할 수 있다.
- [0069] 도 2는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부의 토큰 병합 블록을 설명하는 도면이다.
- [0070] 도 2는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부의 토큰 병합 블록에서 각각 수행되는 절차를 보다 구체적으로 예시 및 설명한다.
- [0071] 도 2를 참고하면, 본 발명의 트랜스포머 인코딩부(200)는 제1 트랜스포머 블록(210), 제2 트랜스포머 블록(220) 및 제3 트랜스포머 블록(230)을 포함한다.
- [0072] 제1 트랜스포머 블록(210), 제2 트랜스포머 블록(220) 및 제3 트랜스포머 블록(230)의 수는 예시로서 증가되거나 감소될 수 있다.
- [0073] 본 발명의 트랜스포머 인코딩부(200)는 제1 트랜스포머 블록(210)에서만 토큰 병합을 진행하여 제1 트랜스포머 블록(210), 제2 트랜스포머 블록(220) 및 제3 트랜스포머 블록(230) 모두에서 토큰 병합을 진행하는 종래 기술과 유사한 정확도 및 개선된 추론 속도를 제공한다.
- [0074] 단계(S201)에서 트랜스포머 인코딩부(200)는 기준 토큰을 선정하고, 선정된 기준 토큰의 인접한 4개의 토큰과 유사도를 기준 유사도에 기반하여 비교한다.
- [0075] 단계(S202)에서 트랜스포머 인코딩부(200)는 인접한 4개의 토큰 중 유사도가 기준 유사도 보다 높은 토큰을 선택한다.
- [0076] 단계(S203)에서 트랜스포머 인코딩부(200)는 선택된 인접 토큰 중 확장 대상으로 기준 토큰을 추가적으로 선택한 이후에 유사도 비교를 단계(S201)와 같이 진행한다.
- [0077] 단계(S204)에서 트랜스포머 인코딩부(200)는 인접한 4개의 토큰 중 유사도가 기준 유사도 보다 높은 토큰을 선택하는 과정을 단계(S202)와 같이 진행한다.
- [0078] 즉, 트랜스포머 인코딩부(200)는 하나의 토큰 기준 상, 하, 좌 및 우 인접 4개 토큰과 유사도 비교를 진행한다.
- [0079] 공간 지역성(spatial locality)과 관련하여 이미지 상에서 인접한 토큰들끼리 비슷한 특징을 가지는 것이 활용된다.
- [0080] 어텐션 분포에 기반하여 선정된 유사도 기준 값이 유사도 비교에 활용된다.
- [0081] 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부의 토큰 병합 블록에서 각각 수행되는 절차는 추가적인 모델 학습 없이 높은 정확도를 유지하면서 전반적인 추론 속도를 개선한다.
- [0082] 본 발명은 적은 횟수로 토큰 유사도 비교 및 공간 지역성을 고려한다.
- [0083] 트랜스포머 인코딩부는 복수의 트랜스포머 블록 중에서 어느 하나에서만 토큰 병합을 진행하여 토큰 병합 소요 시간을 감소시킨다.
- [0084] 바람직하게는 첫 번째 트랜스포머 블록에서 토큰 병합을 진행한다.
- [0086] 도 3 및 도 4는 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법을 설명하는 도면이다.

- [0087] 도 3은 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법이 비전 트랜스포머 가속화를 진행하는 절차를 예시한다.
- [0088] 도 3을 참고하면, 단계(S301)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 입력 데이터를 처리한다.
- [0089] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 인식 대상 이미지를 복수의 패치로 분할하고, 복수의 패치 중 어느 하나의 패치를 토큰 형태의 입력 데이터로 처리할 수 있다.
- [0090] 단계(S302)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 토큰 병합을 수행한다.
- [0091] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 입력 데이터에 대하여 복수의 트랜스포머 블록을 이용하여 트랜스포머 인코딩을 수행하되, 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서만 입력 데이터에 기반한 토큰 매트릭스를 구성하는 복수의 토큰 중 선정된 기준 토큰과 인접 토큰과의 유사도와 기준 유사도에 기반하여 선정된 토큰들에 대한 토큰 병합(token merging)을 수행할 수 있다.
- [0092] 단계(S303)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 토큰 병합 처리 결과에 기반하여 객체 인식 추론을 수행한다.
- [0093] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 복수의 트랜스포머 블록을 통해 출력된 토큰 병합 처리 결과에 기반하여 인식 대상 이미지에서 객체 인식 추론을 수행할 수 있다.
- [0094] 도 4는 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법이 비전 트랜스포머 가속화와 관련된 한번의 토큰 병합(token merging at once, ToMato) 절차를 예시한다.
- [0095] 도 4를 참고하면, 단계(S401)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 토큰 매트릭스를 입력한다.
- [0096] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 인식 대상인 입력 이미지로부터 변환 처리된 토큰 매트릭스를 입력할 수 있다.
- [0097] 단계(S402)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 어텐션 분포를 연산한다.
- [0098] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 어텐션 점수와 소프트 맥스 함수를 적용하여 어텐션 분포를 연산할 수 있다.
- [0099] 단계(S403)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 선택된 기준 토큰에 대해 인접 토큰과 유사도를 비교한다.
- [0100] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 현재 비교 대상인 기준 토큰에 대해 인접한 4개의 토큰과 유사도를 비교한다.
- [0101] 단계(S404)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 비교 토큰의 유사도가 기준 유사도 보다 큰지 작은지 판단한다.
- [0102] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 인접 토큰 중에 기준 유사도보다 큰 유사도를 가지는 토큰이 있는지 판단할 수 있고, 기준 유사도 보다 큰 비교 토큰이 있다면 단계(S405)로 진행하고, 기준 유사도 보다 작다면 단계(S406)으로 진행한다.
- [0103] 단계(S405)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 비교 토큰 정보를 저장하고, 비교 토큰을 기준 토큰으로 선택한다.
- [0104] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 유사도가 기준 유사도보다 큰 경우 토큰들 정보를 저장한다.
- [0105] 단계(S406)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 기 저장된 토큰을 모두 병합한다.
- [0106] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 유사도가 기준 유사도 보다 큰 것

으로 판단되어 저장된 토큰 정보들에 해당하는 모든 토큰을 한번에 저장할 수 있다.

- [0107] 단계(S407)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 전체 토큰에 대해 병합이 완료되었는지 여부를 확인한다.
- [0108] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 기준 유사도 판단 결과에 기반하여 특정 토큰에 대한 병합이 완료되면 다음 토큰으로 넘어가서 새로운 병합 과정을 수행하여 순차적으로 병합 완료 여부를 판단하고, 병합 완료이면 단계(S408)로 진행하고, 병합 완료되지 않으면 단계(S403)로 진행한다.
- [0109] 단계(S408)에서 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 토큰 병합 처리 결과를 출력한다.
- [0110] 즉, 본 발명의 일실시예에 따른 토큰 병합을 이용한 비전 트랜스포머 방법은 토큰 매트릭스 내 토큰에 대한 병합이 완료된 토큰 병합처리 결과를 출력할 수 있다.
- [0111] 따라서, 본 발명은 복수의 트랜스포머 블록 중 어느 하나의 트랜스포머 블록에서 토큰 매트릭스(matrix)를 구성하는 복수의 토큰 중 어느 하나의 토큰을 기준으로 인접한 토큰들과의 유사도를 비교하여 연쇄적으로 토큰 병합을 진행함에 높은 정확도를 유지하면서 전반적인 추론 속도를 개선하는 비전 트랜스포머 장치 및 방법을 제공할 수 있다.
- [0113] 도 5a 내지 도 5e는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서 토큰 병합 절차를 설명하는 도면이다.
- [0114] 도 5a를 참고하면, 토큰 매트릭스(500)와 토큰 매트릭스(500)에 대한 어텐션 분포(501)를 보여준다.
- [0115] 토큰 매트릭스(500)와 관련하여 다중 헤드 어텐션(multi-head attention) 끝난 후에 클래스 토큰과 정제(distillation) 토큰을 제외한 토큰들 간에 어텐션 수행 결과에 대한 분포로 어텐션 분포(501)를 연산한다.
- [0116] 도 5b를 참고하면, 토큰 매트릭스(510)에서 기준 토큰인 i번째 토큰부터 인접한 토큰들 간의 유사도 비교 어텐션 분포(511)에서 수행한다.
- [0117] 어텐션 분포(511)는 실선은 유사하고, 점선은 차이가 존재하는 것을 보여준다.
- [0118] 도 5c를 참고하면, 토큰 매트릭스(520)에서 기준 토큰인 i번째 토큰부터 인접한 토큰들 중 토큰(i+14)을 제외한 인접 토큰을 선택한 결과를 보여준다.
- [0119] 어텐션 분포(521)는 점선에 해당하는 토큰이 토큰 매트릭스(520)에서 선택되지 않음을 보여준다.
- [0120] 즉, 토큰 매트릭스(520) 및 어텐션 분포(521)에서 유사 기준도에 해당하는 $\text{sim}(\text{similarity threshold which affects merging rate})$ 값을 넘는 토큰만 선택됨을 확인할 수 있다.
- [0121] 예를 들어, 유사도 비교와 관련하여 유사도 기준의 값은 $0.1e-7$ 로 설정된 상태일 수 있다.
- [0122] 도 5d를 참고하면, 토큰 매트릭스(530)와 어텐션 분포(531)는 새로 병합된 토큰들을 기준으로 재귀적으로 유사도 비교하는 것을 보여준다.
- [0123] 도 5e를 참고하면, 토큰 매트릭스(540)와 어텐션 분포(541)는 새로 병합된 토큰들을 기준으로 재귀적으로 유사도 비교후 토큰 정보를 저장하는 것을 보여준다.
- [0124] 도 5a 내지 도 5e에서의 이미지 및 표에 따른 토큰 병합 절차를 통합하여 설명한다.
- [0125] 본 발명의 일실시예에 따른 토큰 병합 절차는 첫 번째 디코더(decoder)에서 다중 헤드 어텐션이 진행되고 난 직후 토큰 병합이 시작되는데 상, 하, 좌 및 우로 4개의 인접한 토큰들에 대하여 병합을 진행할 토큰을 판단하는 절차를 진행한다.
- [0126] 병합을 진행할 토큰들을 기준으로 다시 인접한 토큰 4개에 대한 병합을 진행하는 과정을 동일하게 반복한다.
- [0127] 비슷한 토큰들을 재귀적으로 찾아낸 뒤, 더 이상 인접한 토큰들 중 유사한 토큰을 찾을 수 없을 때, 찾아낸 모든 비슷한 토큰을 하나로 병합한다.

- [0128] 재귀적 토큰 병합 과정은 첫번째 토큰에서 시작하여 마지막 토큰까지 진행된다.
- [0130] 도 6은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서 추가 토큰 병합 절차를 설명하는 도면이다.
- [0131] 도 6은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서 추가 토큰 병합 절차와 관련하여 이미지를 이용하여 예시한다.
- [0132] 도 6을 참고하면, 이미지(600)에서 부분(601)에서 인접한 토큰으로 확장하고, 이미지(610)에서 부분(611)과 같이 병합한다.
- [0133] 이미지(620)에서 다음으로 토큰을 넘어가면서, 부분(621)으로 새로운 병합 과정을 진행한다.
- [0134] 인접한 토큰 중 더 이상 유사 기준도를 넘기는 토큰이 없을 경우에 저장된 토큰들을 하나의 토큰으로 병합한 후 다음 토큰으로 넘어가서 반복적으로 병합 과정을 수행한 이후 이미 머징이 완료된 토큰인 경우 다음 부분(621)으로 진행한다.
- [0135] 이미지(620)에서 부분(621)으로 병합 완료를 판단하고, 이미지(630)에서 부분(631)으로 병합하여 전체 토큰에 대한 병합 과정을 수행한다.
- [0136] 전체 토큰에 대한 병합과 관련하여 토큰 병합 처리 결과를 실제 최종 결과를 이미지(640)로 제공한다.
- [0138] 도 7은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서의 토큰 병합 처리 결과를 설명하는 도면이다.
- [0139] 도 7은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 트랜스포머 인코딩부에서의 토큰 병합 처리 결과를 예시한다.
- [0140] 도 7을 참고하면, 입력되는 인식 대상 이미지(700)를 나타내고, 인식 대상 이미지(700)를 본 발명에 적용하여 출력되는 토큰 병합 처리 결과(710)를 나타낸다.
- [0141] 토큰 병합 처리 결과(710)는 의미적으로 패치들을 효과적으로 병합한 결과를 나타낸다.
- [0143] 도 8 및 도 9는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 추론 성능에 대한 시뮬레이션 결과를 설명하는 도면이다.
- [0144] 도 8은 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 추론 성능에 대한 시뮬레이션 결과와 관련하여 종래 기술과 지연속도(latency) 및 정확도(accuracy)를 비교한다.
- [0145] 도 8을 참고하면, 그래프(800)는 본 발명과 관련된 지시선(801)과 종래 기술과 관련된 지시선(802)을 나타낸다.
- [0146] 지시선(801)과 지시선(802)을 비교하면 본 발명이 지연 속도 감소에 따른 정확도 감소가 경미한 것을 확인할 수 있다.
- [0147] 즉, 지시선(802)은 추론 시간에 따른 정확도의 변화폭이 크나, 지시선(801)은 추론 시간이 줄어도 정확도의 변화가 크지 않은 것을 확인할 수 있다.
- [0148] 도 9는 본 발명의 일실시예에 따른 비전 트랜스포머 장치에서 추론 성능에 대한 시뮬레이션 결과와 관련하여 유사 기준도의 값과 추론 속도와 관련된 지연속도(latency)에 대하여 예시한다.
- [0149] 도 9를 참고하면, 그래프(900)는 본 발명과 관련된 지시선(901)으로 유사 기준도의 값과 추론 속도와 관련된 지연속도의 관계를 보여준다.
- [0150] 즉, 그래프(900)는 유사도 기준도 값이 감소할수록 추론 속도가 증가하는 것을 보여준다.
- [0151] 유사도 기준도 값과 관련하여 이미지에 균일한 배경 또는 특징을 가질수록 병합 비율(merging rate)가 높아진다.

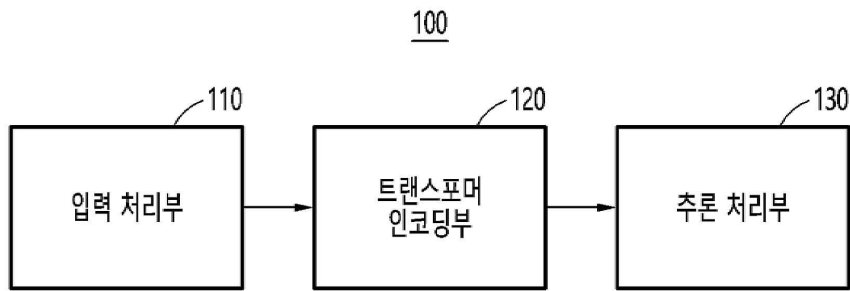
- [0152] 유사도 기준 값과 관련하여 이미지에 중요하거나 복잡한 특징을 많이 가질수록 병합 비율이 낮아진다.
- [0153] 예를 들어, 본 발명은 유사도 기준 값을 10^{-7} 로 설정하여 총 토큰 수를 약 8.16%에서 20.92%로 감소한다.
- [0154] 각 트랜스포머 블록에서 병합되는 토큰들의 수를 나타내는 값을 "2"로 설정했을 때 최종적으로 약 15%의 토큰들이 병합된다.
- [0155] 이에 따라, 본 발명은 유사도 기준 값 설정에 따라 지연 속도와 병합 범위를 유동적으로 설정할 수 있다.
- [0157] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0158] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분포되어서, 분포된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0159] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.
- [0160] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

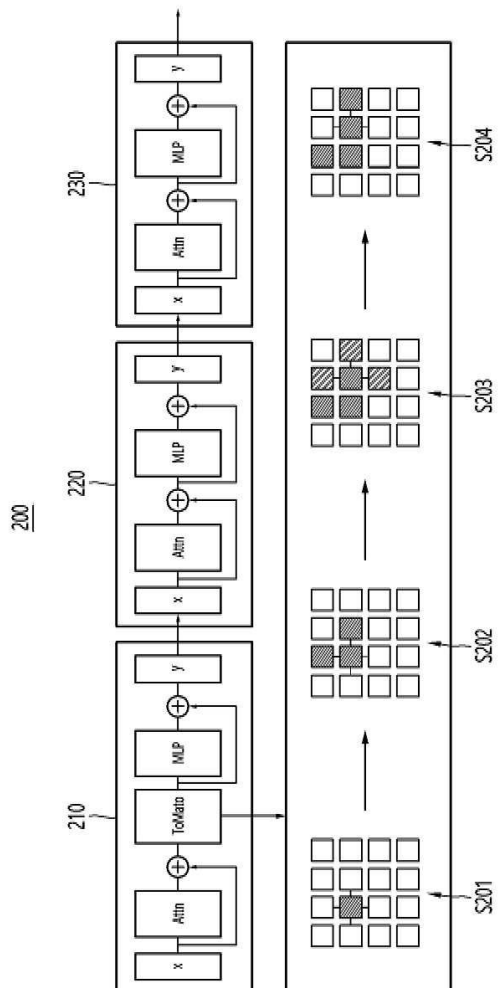
- [0161] 100: 비전 트랜스포머 장치
- 110: 입력 처리부 120: 트랜스포머 인코딩부
- 130: 추론 처리부

도면

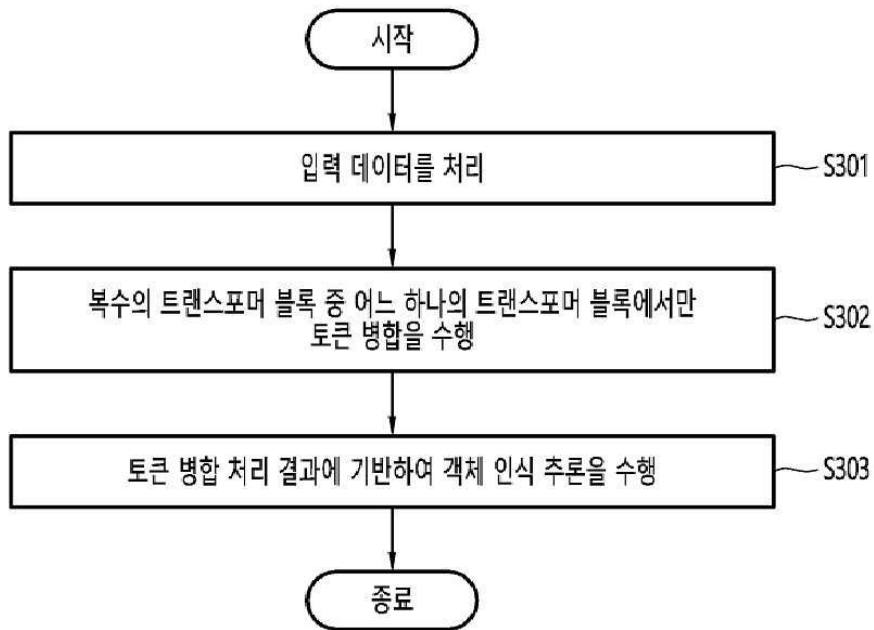
도면1



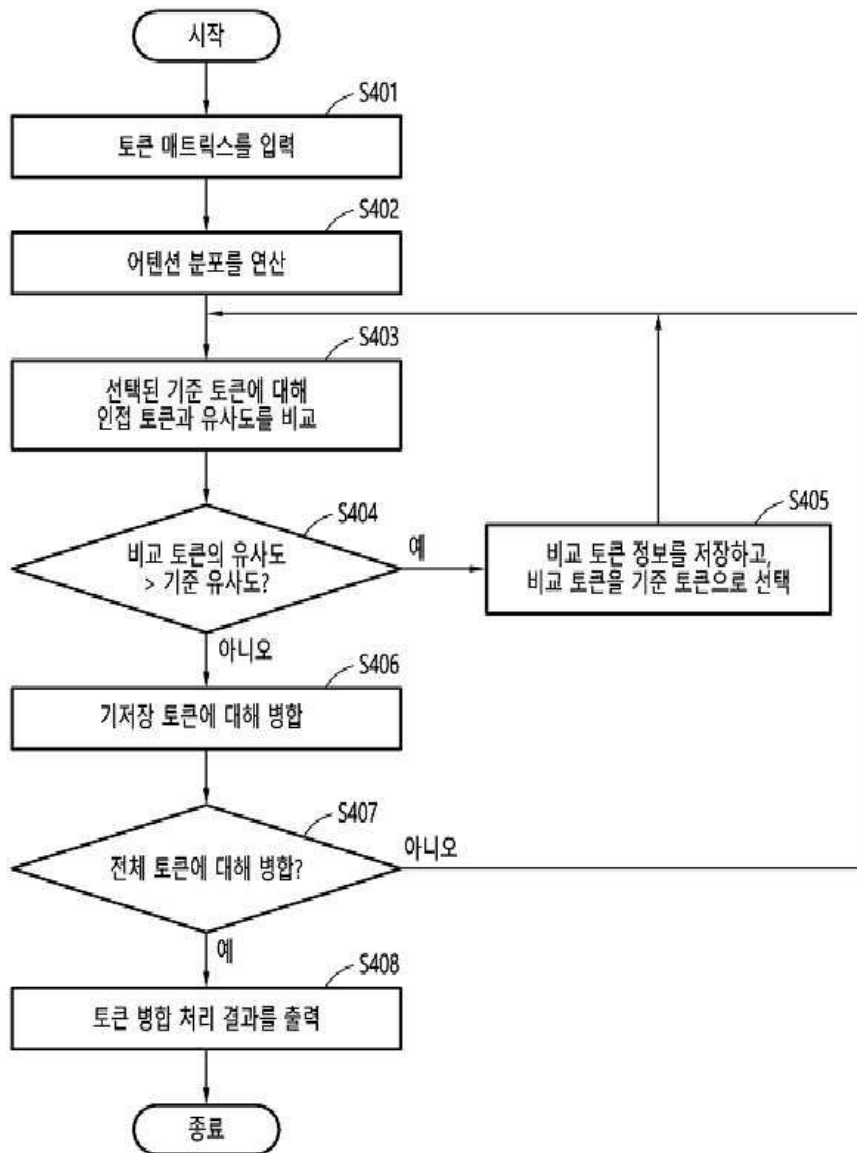
도면2



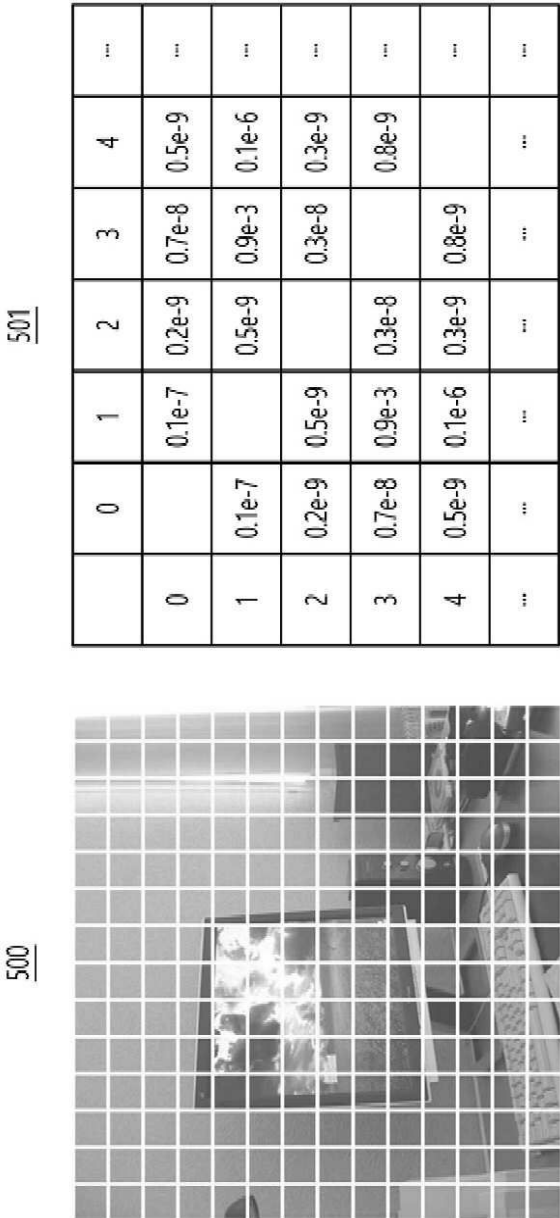
도면3



도면4



도면5a

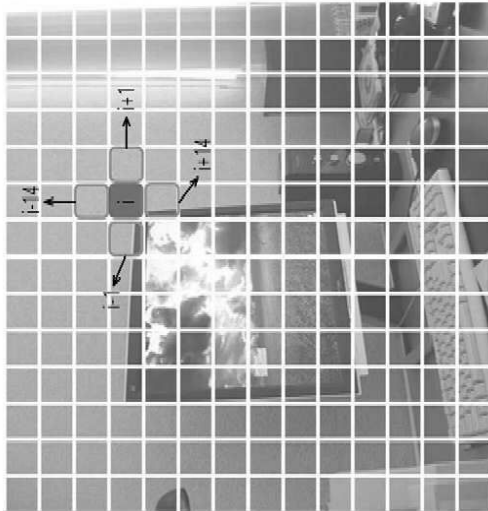


도면5b

511

	i-14	...	i-1	i	i+1	...	i+14
i-2	0.1e-9	...	0.2e-9	0.7e-8	0.5e-9	...	0.1e-9
i-1	0.1e-7	...	0.5e-9	0.9e-3	0.1e-6	...	0.9e-9
i	0.2e-7	...	0.1e-6		0.3e-7	...	0.8e-8
i+1	0.7e-8	...	0.3e-8	0.3e-9	0.8e-9	...	0.4e-7
i+2	0.5e-9	...	0.3e-9	0.8e-9	0.4e-7	...	0.5e-8

510

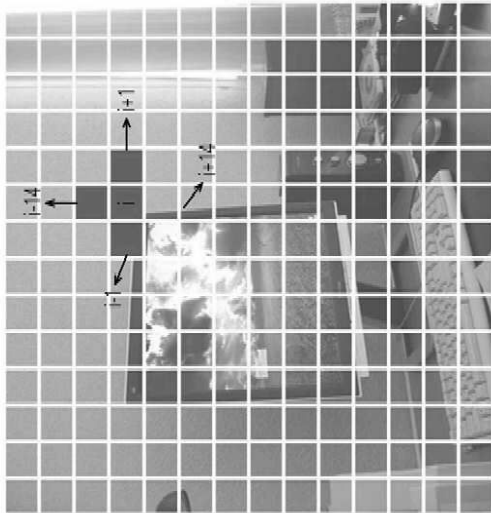


도면5c

521

	i-14	...	i-1	i	i+1	i+14
i-2	0.1e-9	...	0.2e-9	0.7e-8	0.5e-9	0.1e-9
i-1	0.1e-7	...	0.5e-9	0.9e-3	0.1e-6	0.9e-9
i	0.2e-7	...	0.1e-6		0.3e-7	0.8e-8
i+1	0.7e-8	...	0.3e-8	0.3e-9	0.8e-9	0.4e-7
i+2	0.5e-9	...	0.3e-9	0.8e-9	0.4e-7	0.5e-8

520

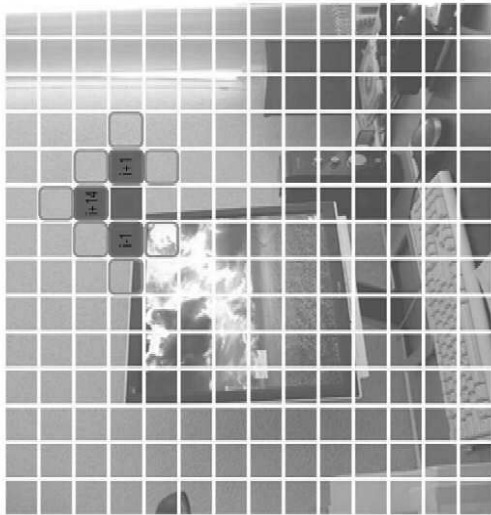


도면5d

531

	i-14	...	i-1	i	i+1	i+14
i-2	0.1e-9	...	0.2e-9	0.7e-8	0.5e-9	0.1e-9
i-1	0.1e-7	...	0.5e-9	0.9e-3	0.1e-6	0.9e-9
i	0.2e-7	...	0.1e-6		0.3e-7	0.8e-8
i+1	0.7e-8	...	0.3e-8	0.3e-9	0.8e-9	0.4e-7
i+2	0.5e-9	...	0.3e-9	0.8e-9	0.4e-7	0.5e-8

530



도면5e

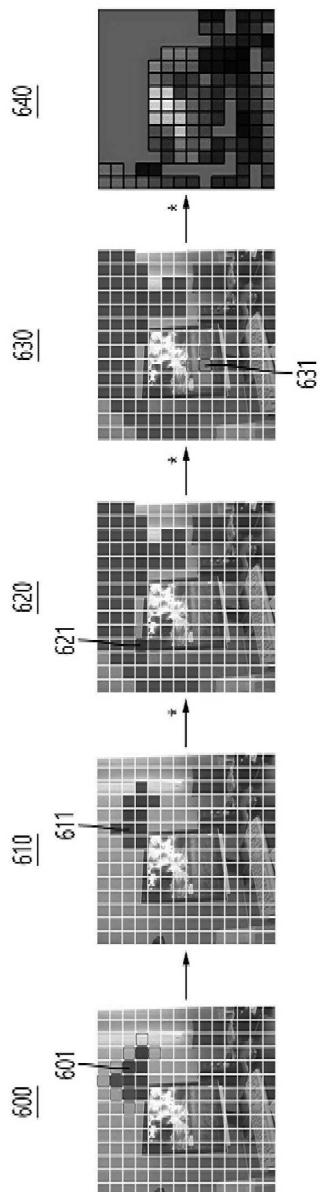
541

	i-14	...	i-1	i	i+1	...	i+14
i-2	0.1e-9	...	0.2e-9	0.7e-8	0.5e-9	...	0.1e-9
i-1	0.1e-7	...	0.5e-9	0.9e-3	0.1e-6	...	0.9e-9
i	0.2e-7	...	0.1e-6		0.3e-7	...	0.8e-8
i+1	0.7e-8	...	0.3e-8	0.3e-9	0.8e-9	...	0.4e-7
i+2	0.5e-9	...	0.3e-9	0.8e-9	0.4e-7	...	0.5e-8

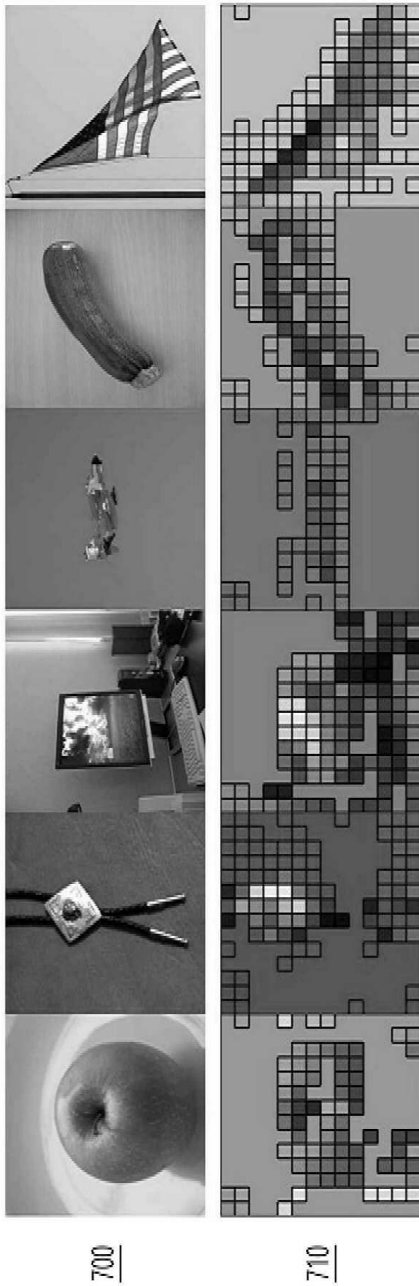
540



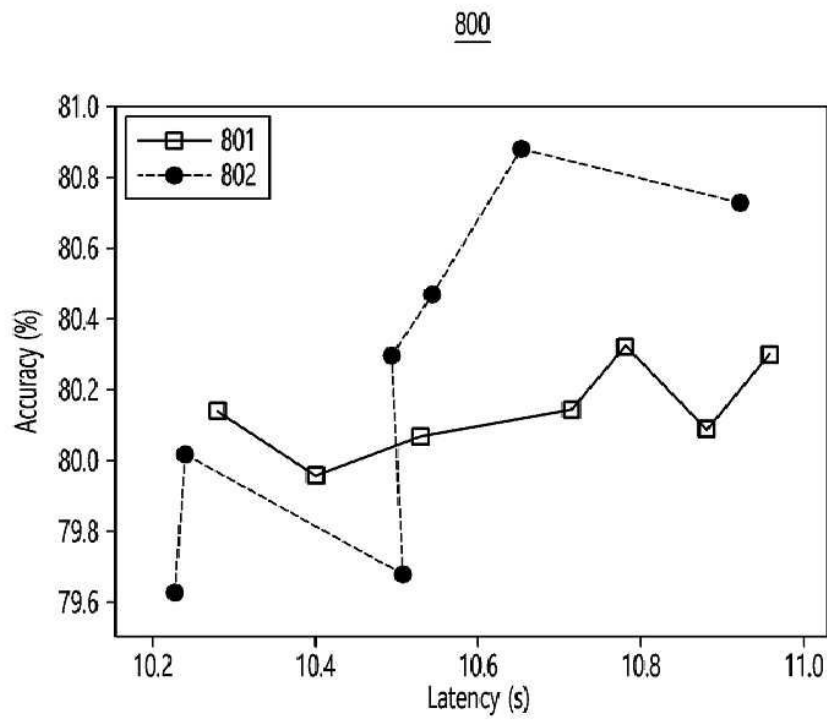
도면6



도면7



도면8



도면9

