(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0135186
(43) 공개일자 2024년09월10일

(51) 국제특허분류(Int. Cl.)
G06N 3/09 (2023.01) *G06N 3/0455* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/09 (2023.01)
G06N 3/0455 (2023.01)
(21) 출원번호 10-2023-0028403
(22) 출원일자 2023년03월03일
심사청구일자 2023년03월03일

(71) 출원인
계명대학교 산학협력단
대구광역시 달서구 달구벌대로 1095, 계명대학교 산학협력관 201호(신당동)
(72) 발명자
고병철
대구광역시 수성구 범어천로 107
김형진
경상북도 청도군 화양읍 한내길 27-3
(74) 대리인
김건우

전체 청구항 수 : 총 16 항

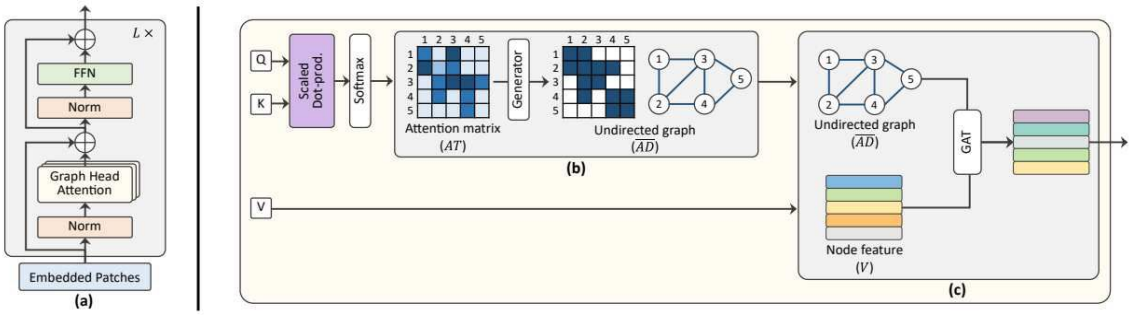
(54) 발명의 명칭 **소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법**

(57) 요 약

본 발명은 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치에 관한 것으로서, 보다 구체적으로는 비전 트랜스포머 장치로서, 비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(Graph Head

(뒷면에 계속)

대 표 도 - 도1



Attention, GHA)을 포함하는 것을 그 구성상의 특징으로 한다.

본 발명은 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치의 동작 방법에 관한 것으로서, 보다 구체적으로는 컴퓨터에서 각 단계가 수행되는 비전 트랜스포머의 동작 방법으로서, (1) 비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(Graph Head Attention, GHA)을 포함하는 비전 트랜스포머인 GHA-ViT를 사전 학습 없이 소규모의 데이터셋으로 학습해 분류 모델을 생성하는 단계; 및 (2) 입력 이미지를 상기 분류 모델에 입력해 분류하는 단계를 포함하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 본 발명의 특징에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법에 따르면, 트랜스포머 헤드 내부에 그래프를 적용하여 MHA를 더 적은 수의 GHA로 대체함으로써, 연산 수를 줄일 수 있고 그래프 구성을 위한 추가 학습 매개변수가 필요하지 않으며, 입력 패치의 지역성과 전역성을 모두 유지하고 어텐션의 다양성을 보장할 수 있으며, 스크래치 훈련만 있고 대규모 데이터셋에 대한 사전 훈련은 없는 중소 데이터셋에 대해 유망한 분류 성능을 얻을 수 있다.

(52) CPC특허분류
G06N 3/082 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345354368
과제번호	2022R1I1A3058128
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공학학술연구기반구축
연구과제명	자기-지도학습 기반의 경량 시공간 그래프 트랜스포머 연구
기 여 율	1/1
과제수행기관명	계명대학교
연구기간	2022.06.01 ~ 2023.02.28

명 세 서

청구범위

청구항 1

비전 트랜스포머 장치로서,

비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(100)(Graph Head Attention, GHA)을 포함하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 2

제1항에 있어서, 상기 그래프 헤드 어텐션(100)은,

입력 패치 사이의 집중도를 나타내는 어텐션 맵을 그래프 구조로 변환하여 사용하되, 상기 멀티-헤드 어텐션보다 적은 수의 헤드를 포함하여 구성되는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 3

제1항에 있어서, 상기 그래프 헤드 어텐션(100)은,

복수의 패치로 구성되는 인코더 계층의 입력으로부터 어텐션 행렬을 계산하고, 어텐션 행렬에 그래프 생성기(graph generator)를 적용해 무방향 그래프를 생성하는 그래프 생성부(110); 및

상기 무방향 그래프를 그래프 네트워크에 적용해 그래프 어텐션을 생성하는 어텐션 생성부(120)를 포함하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 4

제3항에 있어서, 상기 그래프 생성부(110)는,

상기 어텐션 행렬에 Top-k 함수를 적용해 선택된 하위 노드를 그래프 구조로 간주하며, 선택된 하위 노드의 희소 행렬을 인접 행렬(AD)로 보고, 상기 인접 행렬을 사용해 그래프를 구성하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 5

제4항에 있어서, 상기 그래프 생성부(110)는,

상기 인접 행렬의 상삼각 행렬과 상기 상삼각 행렬의 전치행렬을 결합하여 생성한 제1 부분 무방향 그래프와 상기 인접 행렬의 하삼각 행렬과 상기 하삼각 행렬의 전치행렬을 결합하여 생성한 제2 부분 무방향 그래프를 OR 연산하여 상기 무방향 그래프를 생성하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 6

제5항에 있어서, 상기 그래프 생성부(110)는,

훈련 가능한 매개변수인 그래프 강도 및 그래프 가중치를 적용해 상기 무방향 그래프를 보정하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 7

제3항에 있어서, 상기 어텐션 생성부(120)는,

상기 그래프 네트워크로 그래프 어텐션 네트워크(Graph Attention Network, GAT)를 사용하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 8

제7항에 있어서, 상기 어텐션 생성부(120)는,

상기 무방향 그래프와 값(V)에서 추출된 노드 특징을 이용해 상기 그래프 어텐션을 생성하되, 그래프 풀링을 통해 어텐션 점수가 낮은 노드의 링크를 제외하고 인접 노드의 연결을 반영하도록 상기 무방향 그래프에 GAT를 적용하는 GHA 부스팅을 통해, 노드 특징이 업데이트된 그래프 어텐션을 출력하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치.

청구항 9

컴퓨터에서 각 단계가 수행되는 비전 트랜스포머의 동작 방법으로서,

(1) 비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(100)(Graph Head Attention, GHA)을 포함하는 비전 트랜스포머인 GHA-ViT를 사전 학습 없이 소규모의 데이터셋으로 학습해 분류 모델을 생성하는 단계; 및

(2) 입력 이미지를 상기 분류 모델에 입력해 분류하는 단계를 포함하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 10

제9항에 있어서, 상기 그래프 헤드 어텐션(100)은,

입력 패치 사이의 집중도를 나타내는 어텐션 맵을 그래프 구조로 변환하여 사용하되, 상기 멀티-헤드 어텐션보다 적은 수의 헤드를 포함하여 구성되는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 11

제9항에 있어서, 상기 그래프 헤드 어텐션(100)은,

복수의 패치로 구성되는 인코더 계층의 입력으로부터 어텐션 행렬을 계산하고, 어텐션 행렬에 그래프 생성기(graph generator)를 적용해 무방향 그래프를 생성하는 그래프 생성부(110); 및

상기 무방향 그래프를 그래프 네트워크에 적용해 그래프 어텐션을 생성하는 어텐션 생성부(120)를 포함하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 12

제11항에 있어서, 상기 그래프 생성부(110)는,

상기 어텐션 행렬에 Top-k 함수를 적용해 선택된 하위 노드를 그래프 구조로 간주하며, 선택된 하위 노드의 희소 행렬을 인접 행렬(AD)로 보고, 상기 인접 행렬을 사용해 그래프를 구성하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 13

제12항에 있어서, 상기 그래프 생성부(110)는,

상기 인접 행렬의 상삼각 행렬과 상기 상삼각 행렬의 전치행렬을 결합하여 생성한 제1 부분 무방향 그래프와 상기 인접 행렬의 하삼각 행렬과 상기 하삼각 행렬의 전치행렬을 결합하여 생성한 제2 부분 무방향 그래프를 OR 연산하여 상기 무방향 그래프를 생성하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 14

제13항에 있어서, 상기 그래프 생성부(110)는,

훈련 가능한 매개변수인 그래프 강도 및 그래프 가중치를 적용해 상기 무방향 그래프를 보정하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 15

제11항에 있어서, 상기 어텐션 생성부(120)는,

상기 그래프 네트워크로 그래프 어텐션 네트워크(Graph Attention Network, GAT)를 사용하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

청구항 16

제15항에 있어서, 상기 어텐션 생성부(120)는,

상기 무방향 그래프와 값(V)에서 추출된 노드 특징을 이용해 상기 그래프 어텐션을 생성하되, 그래프 풀링을 통해 어텐션 점수가 낮은 노드의 링크를 제외하고 인접 노드의 연결을 반영하도록 상기 무방향 그래프에 GAT를 적용하는 GHA 부스팅을 통해, 노드 특징이 업데이트된 그래프 어텐션을 출력하는 것을 특징으로 하는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 비전 트랜스포머 장치 및 이의 동작 방법에 관한 것으로서, 보다 구체적으로는 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 일실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0004] 트랜스포머는 가장 강력한 신경망 도구 중 하나가 되었으며 자연어 처리(NLP) 및 음성 인식과 같은 시퀀셜 데이터 처리에서 유망한 성능을 보여주었다. 컴퓨터 비전 분야에 적용되는 비전 트랜스포머(Vision Transformer,

ViT)는 이미지 분류, 이미지 분할, 객체 추적, 깊이 추정, 동작 인식 등 다양한 비전 문제에 사용되는 선도적인 알고리즘이다. ViT는 트랜스포머의 전반적인 아키텍처를 크게 바꾸지 않고 이미지 처리에 사용되기 때문에 의미 있는 접근법이다.

[0006] ViT는 일련의 이미지 패치를 입력으로 적용해, 컨볼루션 신경망(CNN)에 의존하지 않는 트랜스포머를 사용하여 CNN 기반 모델의 성능을 능가하는 성능을 달성했다. 이러한 높은 성능에도 불구하고, ViT는 모든 토큰을 동등하게 취급하고 지역성(locality)을 무시하므로, 이미지의 고유한 로컬 구조가 손실되는 문제가 있다. 또한, ViT는 여전히 많은 훈련 데이터와 셀프-어텐션 작업으로 인해 고해상도 이미지를 처리하기 어렵다는 단점이 있다.

[0008] 이러한 문제를 해결하기 위해 트랜스포머 아키텍처를 개선하기 위한 여러 연구가 수행되었다. 이러한 연구는 훈련 데이터의 수와 매개변수를 줄이고 이미지의 위치를 유지하려는 시도로 나눌 수 있다. 그러나 이러한 ViT 기반 방법은 로컬 영역 간 또는 글로벌 및 로컬 영역 간의 공간 기하학적 관계를 제대로 고려하지 않으며 멀티 헤드 어텐션(Multi-Head Attention, MHA) 조합에 크게 의존하기 때문에 계산 수를 줄이는 데 한계가 있다. 또한, ViT에는 많은 훈련 데이터가 필요하므로 사전 훈련 없이 작은 데이터셋으로 훈련하면 성능이 크게 저하된다.

[0010] 트랜스포머에 그래프를 적용한 연구가 다수 있다(Yun et al., 2019; Ying et al., 2021; Yun et al., 2022). 그래프 트랜스포머 네트워크(GTN)(Yun et al., 2019)는 노이즈가 많은 연결을 제외하고 작업(예: metapaths)에 대한 유용한 연결을 포함하는 동시에 새로운 그래프에서 효과적인 노드 표현을 종단 간 방식으로 학습하여 새로운 그래프 구조를 생성하도록 제안되었다. Graphormer(Ying et al., 2021)는 그래프의 구조 정보를 모델로 효과적으로 인코딩하기 위해 그래프를 트랜스포머로 활용하였다. 이 모델에서는 그래프의 노드 중요도를 포착할 수 있는 중심성 인코딩과 노드 간의 구조적 관계를 포착할 수 있는 공간 인코딩이 트랜스포머 인코더 내부에 적용된다. Yun et al.(Yun et al., 2022)는 모델이 트랜스포머 내에서 효과적인 노드를 보존하면서 그래프 노드 간의 노이즈가 많은 연결을 배제하는 또 다른 구조화된 GTN을 제안했다. 3D 포즈 추정을 위해서는 그래프 지향 변환(GraFormer)(Zhao et al., 2022)이 제안되었다. GraFormer는 그래프 어텐션 블록과 그래프 컨볼루션 레이어 블록을 반복적으로 쌓아 그래프 노드의 정보를 융합할 뿐만 아니라 그래프의 위상 구조를 모델링할 수 있다. 그러나 위에서 언급한 모델들은 이미지보다는 노드 분류(예: 인용, 영화, 트래픽, 소셜 네트워킹 및 단백질 상호 작용)와 인간 포즈를 위해 설계되었기 때문에 비전 애플리케이션에 적용하기에는 한계가 있다.

[0012] 일반적인 노드 분류를 위한 그래프 트랜스포머와 달리 비전 애플리케이션을 위한 ViT에 그래프를 적용한 연구는 상대적으로 적다. Shen et al.(Shen et al., 2021)은 차량 재식별을 위한 그래프 대화형 트랜스포머(GiT)를 제안하였다. 이 방법을 통해 GiT는 강력한 글로벌 패치 특징을 추출하기 위한 원래의 트랜스포머 모듈과 패치 내에서 구별되는 로컬 특징을 추출하기 위한 로컬 상관 그래프(local correlation graph, LCG) 모듈의 두 가지 모듈로 나뉜다. LCG 모듈과 트랜스포머 모듈의 출력 특징이 결합되어 다운스트림 작업에 사용된다. Zheng et al.(Zheng et al., 2021)은 전체 슬라이드 영상의 그래프 표현인 그래프-트랜스포머 네트워크와 트랜스포머를 융합하는 방법을 제안하였다. 이미지 패치는 그래프에서 노드 집합이 되고 각 노드는 에지로 연결된다. 생성된 그래프는 그래프 컨볼루션 네트워크(GCN) 및 풀링 레이어를 통해 트랜스포머에 입력된다. 트랜스포머의 출력은 분류를 위해 MLP 헤드에 적용된다. Mesh Graphormer(Lin et al., 2021)는 그래프 컨볼루션을 셀프-어텐션과 통합하여 단일 이미지에서 인간 포즈와 메시를 재구성했다. Mesh Graphormer의 그래프 컨볼루션과 그리드 특징이 모두 순수 트랜스포머의 성능 향상에 도움이 된다는 것을 증명했다.

[0014] 전술한 기존 연구들은 모두 그래프가 어텐션에 직접 적용되지 않고 기존 어텐션 헤드와 그래프가 결합되는 방식이다. 이 경우 특징 조합을 위해 빈번한 특징 차원 스케일링이 발생하며 출력은 전역 특징과 로컬 특징의 진정한 조합이 아니다. 또한 MHA를 그대로 사용하고 별도의 모듈로 그래프를 추가하기 때문에 연산 횟수를 늘리는 한계가 있다.

[0016] 전술한 배경 기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지 기술이라 할 수는 없다.

선행기술문헌

비특허문헌

[0017] (비특허문헌 0001) (0001) Shen, F., Xie, Y., Zhu, J., Zhu, X., and Zeng, H. Git: Graph interactive transformer for vehicle re-identification. arXiv preprint arXiv:2107.05475, 2021.

(비특허문헌 0002) (0002) Zheng, Y., Gindra, R., Betke, M., Beane, J. E., and Kolachalama, V. B. A deep learning based graph-transformer for whole slide image classification. medRxiv, 41:1-14, 2021.

(비특허문헌 0003) (0003) Lin, K., Wang, L., and Liu, Z. Mesh graphormer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12939-12948, 2021.

(비특허문헌 0004) (0004) Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICML), pp. 1-21, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

(비특허문헌 0005) (0005) Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., and Shi, H. Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704, 2021.

ČĆ(비특허문헌 0006) (0006) Velikovi, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.

발명의 내용

해결하려는 과제

[0018] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 트랜스포머 헤드 내부에 그래프를 적용하여 MHA를 더 적은 수의 그래프 헤드 어텐션(Graph Head Attention, GHA)으로 대체함으로써, 연산 수를 줄일 수 있고 그래프 구성을 위한 추가 학습 매개변수가 필요하지 않으며, 입력 패치의 지역성과 전역성을 모두 유지하고 어텐션의 다양성을 보장할 수 있으며, 스크래치 훈련만 있고 대규모 데이터셋에 대한 사전 훈련은 없는 중소 데이터셋에 대해 유망한 분류 성능을 보이는, 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법을 제공하는 것을 그 목적으로 한다.

[0020] 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있고, 명시적으로 언급하지 않더라도 과제의 해결수단이나 실시 형태로부터 파악될 수 있는 목적이나 효과도 이에 포함됨은 물론이다.

과제의 해결 수단

[0021] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치는,

[0022] 비전 트랜스포머 장치로서,

[0023] 비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(Graph Head Attention, GHA)을 포함하는 것을 그 구성상의 특징으로 한다.

[0025] 바람직하게는, 상기 그래프 헤드 어텐션은,

[0026] 입력 패치 사이의 집중도를 나타내는 어텐션 맵을 그래프 구조로 변환하여 사용하되, 상기 멀티-헤드 어텐션보다 적은 수의 헤드를 포함하여 구성될 수 있다.

[0028] 바람직하게는, 상기 그래프 헤드 어텐션은,

[0029] 복수의 패치로 구성되는 인코더 계층의 입력으로부터 어텐션 행렬을 계산하고, 어텐션 행렬에 그래프 생성기(graph generator)를 적용해 무방향 그래프를 생성하는 그래프 생성부; 및

[0030] 상기 무방향 그래프를 그래프 네트워크에 적용해 그래프 어텐션을 생성하는 어텐션 생성부를 포함할 수 있다.

[0032] 더욱 바람직하게는, 상기 그래프 생성부는,

[0033] 상기 어텐션 행렬에 Top-k 함수를 적용해 선택된 하위 노드를 그래프 구조로 간주하며, 선택된 하위 노드의 희소 행렬을 인접 행렬(AD)로 보고, 상기 인접 행렬을 사용해 그래프를 구성할 수 있다.

[0035] 더더욱 바람직하게는, 상기 그래프 생성부는,

[0036] 상기 인접 행렬의 상삼각 행렬과 상기 상삼각 행렬의 전치행렬을 결합하여 생성한 제1 부분 무방향 그래프와 상기 인접 행렬의 하삼각 행렬과 상기 하삼각 행렬의 전치행렬을 결합하여 생성한 제2 부분 무방향 그래프를 OR 연산하여 상기 무방향 그래프를 생성할 수 있다.

[0038] 더욱더 바람직하게는, 상기 그래프 생성부는,

[0039] 훈련 가능한 매개변수인 그래프 강도 및 그래프 가중치를 적용해 상기 무방향 그래프를 보정할 수 있다.

[0041] 더욱 바람직하게는, 상기 어텐션 생성부는,

[0042] 상기 그래프 네트워크로 그래프 어텐션 네트워크(Graph Attention Network, GAT)를 사용할 수 있다.

[0044] 더더욱 바람직하게는, 상기 어텐션 생성부는,

[0045] 상기 무방향 그래프와 값(V)에서 추출된 노드 특징을 이용해 상기 그래프 어텐션을 생성하되, 그래프 풀링을 통해 어텐션 점수가 낮은 노드의 링크를 제외하고 인접 노드의 연결을 반영하도록 상기 무방향 그래프에 GAT를 적용하는 GHA 부스팅을 통해, 노드 특징이 업데이트된 그래프 어텐션을 출력할 수 있다.

[0047] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치의 동작 방법은,

[0048] 컴퓨터에서 각 단계가 수행되는 비전 트랜스포머의 동작 방법으로서,

[0049] (1) 비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(Graph Head Attention, GHA)을 포함하는 비전 트랜스포머인 GHA-ViT를 사전 학습 없이 소규모의 데이터셋으로 학습해 분류 모델을 생성하는 단계; 및

[0050] (2) 입력 이미지를 상기 분류 모델에 입력해 분류하는 단계를 포함하는 것을 그 구성상의 특징으로 한다.

[0052] 바람직하게는, 상기 그래프 헤드 어텐션은,

[0053] 입력 패치 사이의 집중도를 나타내는 어텐션 맵을 그래프 구조로 변환하여 사용하되, 상기 멀티-헤드 어텐션보

다 적은 수의 헤드를 포함하여 구성될 수 있다.

[0055] 바람직하게는, 상기 그래프 헤드 어텐션은,

[0056] 복수의 패치로 구성되는 인코더 계층의 입력으로부터 어텐션 행렬을 계산하고, 어텐션 행렬에 그래프 생성기(graph generator)를 적용해 무방향 그래프를 생성하는 그래프 생성부; 및

[0057] 상기 무방향 그래프를 그래프 네트워크에 적용해 그래프 어텐션을 생성하는 어텐션 생성부를 포함할 수 있다.

[0059] 더욱 바람직하게는, 상기 그래프 생성부는,

[0060] 상기 어텐션 행렬에 Top-k 함수를 적용해 선택된 하위 노드를 그래프 구조로 간주하며, 선택된 하위 노드의 희소 행렬을 인접 행렬(AD)로 보고, 상기 인접 행렬을 사용해 그래프를 구성할 수 있다.

[0062] 더더욱 바람직하게는, 상기 그래프 생성부는,

[0063] 상기 인접 행렬의 상삼각 행렬과 상기 상삼각 행렬의 전치행렬을 결합하여 생성한 제1 부분 무방향 그래프와 상기 인접 행렬의 하삼각 행렬과 상기 하삼각 행렬의 전치행렬을 결합하여 생성한 제2 부분 무방향 그래프를 OR 연산하여 상기 무방향 그래프를 생성할 수 있다.

[0065] 더욱더 바람직하게는, 상기 그래프 생성부는,

[0066] 훈련 가능한 매개변수인 그래프 강도 및 그래프 가중치를 적용해 상기 무방향 그래프를 보정할 수 있다.

[0068] 더욱 바람직하게는, 상기 어텐션 생성부는,

[0069] 상기 그래프 네트워크로 그래프 어텐션 네트워크(Graph Attention Network, GAT)를 사용할 수 있다.

[0071] 더더욱 바람직하게는, 상기 어텐션 생성부는,

[0072] 상기 무방향 그래프와 값(V)에서 추출된 노드 특징을 이용해 상기 그래프 어텐션을 생성하되, 그래프 풀링을 통해 어텐션 점수가 낮은 노드의 링크를 제외하고 인접 노드의 연결을 반영하도록 상기 무방향 그래프에 GAT를 적용하는 GHA 부스팅을 통해, 노드 특징이 업데이트된 그래프 어텐션을 출력할 수 있다.

발명의 효과

[0073] 본 발명에서 제안하고 있는 본 발명의 특징에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법에 따르면, 트랜스포머 헤드 내부에 그래프를 적용하여 MHA를 더 적은 수의 GHA로 대체함으로써, 연산 수를 줄일 수 있고 그래프 구성을 위한 추가 학습 매개변수가 필요하지 않으며, 입력 패치의 지역성과 전역성을 모두 유지하고 어텐션의 다양성을 보장할 수 있으며, 스크래치 훈련만 있고 대규모 데이터셋에 대한 사전 훈련은 없는 중소 데이터셋에 대해 유망한 분류 성능을 얻을 수 있다.

[0075] 더불어, 본 발명의 다양하면서도 유익한 장점과 효과는 상술한 내용에 한정되지 않으며, 본 발명의 구체적인 실시 형태를 설명하는 과정에서 보다 쉽게 이해될 수 있을 것이다.

도면의 간단한 설명

[0076] 도 1은 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치의 전체 구조를 나타낸 도면.

도 2는 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치

에서, 그래프 헤드 어텐션의 구성을 도시한 도면.

도 3은 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치에서, 어텐션 행렬 AT로부터의 무방향 그래프를 생성하는 과정을 나타낸 도면.

도 4는 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치의 동작 방법의 흐름을 도시한 도면.

도 5는 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치에서, 실험에 사용된 기준 GHA-ViT 모델의 변형을 요약하여 도시한 도면.

도 6은 소규모 데이터셋에서 스크래치 훈련된 CNN과 트랜스포머 기반 모델의 성능을 비교하여 나타낸 도면.

도 7은 ImageNet-1K 데이터셋에 대한 본 발명의 GHA-ViT 방법과 SoTA 방법의 성능 비교 결과를 나타낸 도면.

도 8은 그래프 구조 생성 방법에 따른 정확도(%)를 비교하여 나타낸 도면.

도 9는 다중-판독 특징의 효과를 입증하기 위해 소규모 데이터셋에서 서로 다른 판독 계층의 정확도(%)를 비교해 나타낸 도면.

도 10은 소규모 데이터셋에 대한 GHA-S에 여러 유형의 그래프 컨볼루션을 적용하였을 때의 정확도(%)를 비교하여 나타낸 도면.

도 11은 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법에서 생성된 그래프 구조를 시각화하여 나타낸 도면.

발명을 실시하기 위한 구체적인 내용

[0077] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0079] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0081] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.

[0083] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.

[0085] 이하, 첨부된 도면을 참고하여 본 발명의 실시예들을 상세히 설명하도록 한다.

[0087] 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머(GHA-ViT) 장치 및 이의 동작 방법은, ViT의 연산 수를 줄이고 이미지 분류를 위한 글로벌 및 로컬 특징을 보존하기 위해서, 그래프 생성 및 그래프 어텐션을 사용하여 더 적은 그래프 헤드로 멀티 헤드 어텐션(MHA)를 대체하는 ViT용 그

래프 헤드 어텐션(GHA)을 포함할 수 있다.

[0089] 본 발명은 i) 그래프와 어텐션을 병렬로 동작시키고 출력을 조합하는 다른 그래프 기반 트랜스포머(Shen et al., 2021; Zheng et al., 2021; Lin et al., 2021)와 달리 트랜스포머 헤드 내부에 그래프를 적용하여 MHA를 더 적은 수의 GHA로 대체하려는 최초의 시도이다. ii) 순수한 ViT와는 달리 패치 임베딩에서 클래스 토큰이 필요 없고, 따라서 연산 수를 줄일 수 있다. iii) 그래프 구조는 어텐션 행렬을 사용하여 구성되고 노드 특징은 값(V) 행렬에서 추출되므로 그래프 구성을 위한 추가 학습 매개변수가 필요하지 않다. iv) 그래프 풀링을 통해 어텐션 점수가 낮은 노드의 링크를 제외하고, 인접 노드의 연결을 반영하도록 GHA 부스팅을 적용하여 노드 특징을 업데이트할 수 있다. 이 프로세스를 통해 특징의 지역성을 보존하고 어텐션의 다양성을 확보할 수 있다. v) GHA-ViT는 스크래치 훈련만 있고 사전 훈련은 없는 중소 데이터셋에 대해서 유망한 분류 성능을 보여준다.

[0091] 이하에서는, 먼저 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치 및 이의 동작 방법을 설명하기 위해, 비전 트랜스포머, 패치 생성 및 그래프 어텐션 네트워크에 대해 각각 설명하도록 한다.

[0093] 비전 트랜스포머(ViT)

[0094] ViT는 최소한의 수정으로 표준 트랜스포머를 이미지에 직접 적용한다. 이를 위해 ViT는 이미지를 특징 차원이 d인 P 패치(예: NLP의 토큰)로 분할하고 트랜스포머에 입력으로 사용하며, 이러한 패치의 선형 임베딩 시퀀스를 제공한다. 임베디드 패치 행렬 $H \in R^{Pxd}$가 정규화된 후, MHA 모듈을 통과하여 MLP로 공급되어 인코더의 출력을 생성한다. 인코더는 L층으로 구성되며 1부터 L까지 반복적으로 작동한다. 마지막으로, 예측 헤드 가중치 $W_c \in R^{dx|C|}$와 바이어스 $b_c \in R^{|C|}$를 사용한 계산을 통해 표현 y를 구한다. 여기서 |C|는 클래스의 수이다.

[0096] 패치 생성

[0097] 순수한 ViT는 단일 컨볼루션 레이어 기반 이미지 토큰화기(image tokenizer)를 적용하여 주어진 이미지를 작은 패치에 배치한다. 이미지를 주어진 패치 크기에 따라 여러 패치로 분할한다. 이와 관련하여 각 패치는 이미지의 위치를 보호하는 데 근본적인 어려움이 있다. 따라서 유도 편향을 학습하려면 상당한 수의 데이터셋을 학습해야 한다. 본 발명에서는 이를 극복하기 위해 풀링 레이어 기반 토큰화기(Hassani et al., 2021)를 사용한다. 유도 편향을 강화하는 데 도움이 되는 풀링 레이어를 사용하여 빅 데이터 패러다임을 벗어나는 데 도움이 된다.

[0099] 그래프 어텐션 네트워크(GAT)

[0100] GAT(Veli$\check{C}$kovi$\acute{C}$ et al., 2017)는 마스크된 셀프-어텐션 계층을 사용하여 그래프 구조 데이터에 적용할 수 있는 컨볼루션 스타일의 신경망이다. GCN(Kipf & Welling, 2016; Lee et al., 2021)과 달리 GAT는 동일한 인접 그룹의 노드에 다른 중요도를 할당하므로 모델 용량을 향상시키고 해석에 도움이 될 수 있다. 전체 그래프에 접근하지 않고도 이러한 학습이 가능하기 때문에 귀납적 학습이 가능한 구조를 가지고 있다. 어텐션 점수는 다음 수학식 1과 같이 노드 i의 이웃의 중요도를 결정하여 입력 데이터를 재정의한다:

수학식 1

[0102] $$h_i' = \sigma\left(\sum_{j \in O_i} \alpha_{ij} h_j W\right)$$

[0104] 여기서 O는 노드 번호, α는 정규화된 어텐션 점수, W는 노드 특징 h에 대한 훈련 가능한 매개변수이다.

[0106] 도 1은 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치의 전체 구조를 나타낸 도면이다. 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션 기반의 비전 트랜스포머 장치는, 비전 트랜스포머에 있어서, 인코더 계층의 어텐션 헤드 내부에 그래프를 적용하여 멀티-헤드 어텐션(Multi-Head Attention, MHA)을 대체하는 그래프 헤드 어텐션(Graph Head Attention, GHA)을 포함할 수 있다. 보다 구체적으로, 그래프 헤드 어텐션은, 입력 패치 사이의 집중도를 나타내는 어텐션 맵을 그래프 구조로 변환하여 사용하되, 멀티-헤드 어텐션보다 적은 수의 헤드를 포함하여 구성될 수 있다.

[0108] 도 2는 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치에서, 그래프 헤드 어텐션(100)의 구성을 도시한 도면이다. 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 그래프 헤드 어텐션(100)은, 복수의 패치로 구성되는 인코더 계층의 입력으로부터 어텐션 행렬을 계산하고, 어텐션 행렬에 그래프 생성기(graph generator)를 적용해 무방향 그래프를 생성하는 그래프 생성부(110); 및 무방향 그래프를 그래프 네트워크에 적용해 그래프 어텐션을 생성하는 어텐션 생성부(120)를 포함하여 구성될 수 있다.

[0110] 도 1의 (a)는 몇 개의 그래프 헤드 어텐션(100)을 포함하여 구성된 GHA 인코더 계층을 도시한 것이고, (b)는 그래프 생성부(110)의 무방향 그래프 생성 과정을 나타낸 것이며, (c)는 어텐션 생성부(120)의 최종 GHA 생성 과정을 나타낸 것이다. 보다 구체적으로, 도 1의 (b)에 도시된 바와 같이, 어텐션 점수는 쿼리(Q)와 전치 키(K)의 스케일 점 연산(scaled dot product)으로 계산될 수 있다. 그 다음, 하위 노드를 선택하기 위해서, 어텐션 행렬 AT에 그래프 생성기를 적용하여 무방향 그래프(undirected graph)로 변환할 수 있다. 도 1의 (c)에 도시된 바와 같이, 무방향 그래프 생성 후에는 값(V)과 무방향 그래프를 GAT(Graph Attention Network)에 적용해 그래프 어텐션을 생성할 수 있다. 생성된 그래프 및 노드 특징 V에 기초하여, GAT는 인접 노드에 다른 중요도를 제공할 수 있다.

[0112] 이하에서는, 도 1 및 도 2를 참조하여 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치에서, 그래프 헤드 어텐션(100)의 구성에 대해 상세히 설명하도록 한다.

[0114] 그래프 생성부(110)는, 복수의 패치로 구성되는 인코더 계층의 입력으로부터 어텐션 행렬을 계산하고, 어텐션 행렬에 그래프 생성기(graph generator)를 적용해 무방향 그래프를 생성할 수 있다. 어텐션 헤드는 트랜스포머의 핵심 아이디어이며, 쿼리(Q), 키(K) 및 값(V)의 세 가지 행렬을 사용하여 각 패치 사이의 어텐션을 계산할 수 있다. 어텐션 점수는 Q와 전이된 K의 스케일 점 연산(scaled dot product)으로 계산되며, 병렬로 실행할 수 있다. 따라서 계산 속도 측면에서 이전 신경망보다 유리하다.

[0116] 트랜스포머는 MHA를 사용하므로, n개의 헤드는 입력으로부터 서로 다른 어텐션을 학습하고 이들의 조합을 통해 강력한 어텐션 표현을 얻을 수 있다(Dosovitsky et al., 2021). 그러나 현실적으로 MHA의 모든 헤드가 트랜스포머의 어텐션 성능에 동일한 영향을 미치는 것은 아니다. 오히려, 헤드의 일부만이 트랜스포머의 어텐션 성능에 영향을 미치고, 나머지는 불필요한 부분에 집중하여 최종 어텐션의 결과에 부정적인 영향을 미친다(Michel et al., 2019). 이러한 관점에서, MHA가 트랜스포머에서 필수적인 것은 아니라는 것은 분명하다. 본 발명에서는 여러 헤드를 사용하지 않고 더 적은 GHA들로부터 더 나은 어텐션을 받을 수 있는 새로운 트랜스포머를 제안한다.

[0118] X를 인코더 계층의 입력이라고 하자. 여기서 X는 P개의 패치로 구성되며, 입력 패치의 숨겨진 차원은 d_h이다. 그러면 i번째 패치는 $x_i \in \mathbb{R}^{d_h}$ 및 $X = [x_1, x_2, \cdots, x_p]$, $X \in \mathbb{R}^{P \times d_h}$로 표시될 수 있다. Q, K, V 행렬은 각각 대응하는 가중치 행렬 W_Q, W_K, $W_V \in \mathbb{R}^{d \times d_h}$을 가지며, Q, K 및 V는 입력 X와 가중치 행렬의 점 연산(dot product)을 통해 얻을 수 있다. 헤드의 어텐션 행렬(AT)은 다음 수학식 2 및 3과 같이 계산될 수 있다:

수학식 2

[0120]
$$\begin{aligned} Q &= XW_Q \\ K &= XW_K \\ V &= XW_V \end{aligned} \quad Q, K, V \in \mathbb{R}^{P \times d_h}$$

수학식 3

[0122]
$$AT = \text{softmax}(\frac{QK^T}{\sqrt{d_h}}) \in \mathbb{R}^{P \times P}$$

[0124] 그래프 생성부(110)는, 어텐션 행렬에 Top-k 함수를 적용해 선택된 하위 노드를 그래프 구조로 간주하며, 선택된 하위 노드의 희소 행렬을 인접 행렬(AD)로 보고, 인접 행렬을 사용해 그래프를 구성할 수 있다. 보다 구체적으로, 그래프 생성부(110)는, 인접 행렬의 상삼각 행렬(upper matrix)과 상삼각 행렬의 전치행렬을 결합하여 생성한 제1 부분 무방향 그래프와 인접 행렬의 하삼각 행렬(lower matrix)과 하삼각 행렬의 전치행렬을 결합하여 생성한 제2 부분 무방향 그래프를 OR 연산하여 무방향 그래프를 생성할 수 있다. 또한, 그래프 생성부(110)는, 훈련 가능한 매개변수인 그래프 강도 및 그래프 가중치를 적용해 무방향 그래프를 보정할 수 있다.

[0126] 도 3은 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치에서, 어텐션 행렬 AT로부터 무방향 그래프를 생성하는 과정을 나타낸 도면이다. 여기서, 도 3의 (a)는 Top-k 함수를 이용해 임계값을 적용해 생성한 초기 인접 행렬 및 그래프, (b)는 상삼각 행렬 및 제1 부분 무방향 그래프, (c)는 하삼각 행렬 및 제2 부분 무방향 그래프, (d)는 무방향 그래프를 각각 나타낸다. 여기서, ∨는 OR 연산이다. 이하에서는, 도 3을 참조하여 그래프 생성부(110)가 무방향 그래프를 구성하는 과정을 상세히 설명하도록 한다.

[0128] CNN에서 더 나은 일반화와 성능을 제공하기 위해 풀링 계층(pooling layer)은 특징 맵의 크기를 줄이고 수용 영역을 넓히는 데 중요한 역할을 한다. 그러나 그래프 노드 간에는 로컬 정보가 없기 때문에 이 풀링 작업을 그래프에 직접 적용할 수는 없다. 따라서 본 발명에서는, 마스크 필터(mask filter)를 기반으로 한 로컬 정보를 포함하는 그래프 풀링(graph pooling)을 제안한다. 그래프 풀링은 그래프 데이터의 다운샘플링을 허용하고 노드의 하위 집합을 적응적으로 선택하여 새롭고 작은 그래프를 형성할 수 있다.

[0130] 이와 관련하여, 본 발명에서는 유의한 연결성을 가진 하위 노드를 선택하기 위해 어텐션 행렬 AT에 Top-k 함수를 적용할 수 있다. 증류된 희소 노드(sparse nodes)는 어텐션 행렬에서 파생된 새로운 형태의 그래프 구조로 간주될 수 있다. 즉, 이러한 노드의 희소 행렬을 그래프의 인접 행렬(adjacency matrix) AD로 간주할 수 있다. 전체 패치 수 P 및 전체 레이어 수 L을 갖는 구조의 경우, 각 층에 대한 k 값은 전체 패치 수의 최소 10%에서 최대 80%까지 (0.8×P)/L씩 증가하면서 선택되도록 설정될 수 있다.

수학식 4

[0132] $$AD = Top - k(AT, k)$$

[0134] 이제 AD를 사용하여 노드 패치로 구성된 그래프를 구성할 수 있다.

[0136] 다시, 노드의 자체-에지(self-edge)를 고려하기 위해 AD에 단위 행렬(identity matix) I를 추가할 수 있다. 다만, 도 3의 (a)에 나타낸 것과 같이 AD로부터 구성된 초기 그래프에서는, 방향(directed) 에지와 무방향(undirected) 에지가 혼합되어 있다. 트랜스포머에서 어텐션은 인접 패치와의 상호 작용에 의해 생성되기 때문에 방향 에지는 올바른 패치 어텐션을 보장할 수 없다. 따라서 혼합된 그래프를 무방향 그래프로 변환해야 한다. 이를 위해 다음과 같은 그래프 변환 방법을 사용할 수 있다.

[0138] 먼저, 도 3의 (b) 및 아래 수학식 5와 같이 AD의 상삼각 행렬과 그 전치(transpose)를 첨가하여 부분적으로 무방향인 그래프(제1 부분 무방향 그래프)를 형성할 수 있다.

수학식 5

[0140] $$AD_{triU} = triU(AD) + triU(AD)^T$$

[0142] 마찬가지로, 도 3의 (c) 및 아래 수학식 6과 같이 AD의 하삼각 행렬과 그 전치도 부분적으로 무방향인 그래프(제2 부분 무방향 그래프)를 형성하기 위해 추가될 수 있다.

수학식 6

[0144] $$AD_{triL} = triL(AD) + triL(AD)^T$$

[0146] 마지막으로, 도 3의 (d)와 같이 제1 부분 무방향 행렬 AD_{triU}과 제1 부분 무방향 행렬 AD_{triL}은 OR 연산($\vee$)을 통해 무방향 그래프 $\overline{AD}$ 를 생성할 수 있다.

수학식 7

[0148] $$\overline{AD} = AD_{triU} \vee AD_{triL}$$

[0150] 또한, 불규칙한 연결에도 강건한 그래프를 만들기 위해 그래프 보정 프로세스를 적용할 수 있다. 그래프 강도

$\overline{AD}_I$ 및 그래프 가중치 $\overline{AD}_w$를 다음 수학식 8과 같이 적용할 수 있다.

수학식 8

[0152] $$\overline{AD} := (\overline{AD} + \overline{AD}_I) \odot \overline{AD}_w$$

[0154] 여기서, $\overline{AD}_I, \overline{AD}_w \in \mathbb{R}^{P \times P}$ 는 훈련 가능한 매개변수이고, ⊙는 Hadamard 연산을 나타낸다.

[0156] 어텐션 생성부(120)는, 무방향 그래프를 그래프 네트워크에 적용해 그래프 어텐션을 생성할 수 있다. 보다 구체적으로, 어텐션 생성부(120)는, 그래프 네트워크로 그래프 어텐션 네트워크(Graph Attention Network, GAT)를 사용할 수 있다. 이때, 어텐션 생성부(120)는, 무방향 그래프와 값(V)에서 추출된 노드 특징을 이용해 그래프 어텐션을 생성하되, 그래프 풀링을 통해 어텐션 점수가 낮은 노드의 링크를 제외하고 인접 노드의 연결을 반영하도록 무방향 그래프에 GAT를 적용하는 GHA 부스팅을 통해, 노드 특징이 업데이트된 그래프 어텐션을 출력할 수 있다.

[0158] 이미지 분류의 정확도를 향상시키기 위해 MHA의 조합을 통해 자세한 어텐션을 얻을 수 있다. 그러나 MHA에는 각 헤드에 대한 가중치 행렬 W_Q, W_K 및 W_V가 필요하다. 따라서 헤드 수가 증가함에 따라 더 많은 학습 매개변수, 메모리 및 계산 시간이 필요하다. 또한, 헤드 수가 일정 기준을 초과하면 각 헤드의 어텐션이 유사해지기 때문에, 테스트 중에 많은 어텐션 헤드를 제거해도 성능에 큰 영향을 미치지 않는다(Michel et al., 2019).

[0160] 본 발명에서는, MHA로 인한 문제를 피하기 위해, 적은 수의 GHA를 사용하고, 제안된 GHA만 사용하여 트랜스포머가 성공적으로 작동할 수 있음을 보여준다. 어텐션의 다양성을 보장하고, 그래프 노드 간의 상관관계를 강조하기 위해, GHA에 GAT(VeliČkoviĆ et al., 2017)를 적용할 수 있다. 일반적인 MHA 메커니즘 대신 GHA를 사용할 때 어텐션의 다양성을 확보하는 것은 GHA 모델의 성공적인 작동에 매우 중요한 부분이다.

[0162] 그래프 생성부(110)에서는, 어텐션 행렬에서 노드 패치 간의 관계를 나타내는 $\overline{AD}$를 추출했다. 어텐션 생성부(120)에서는, 추출된 $\overline{AD}$ 가 있는 노드 사이에서 효율적인 어텐션 계산을 수행하기 위해 $\overline{AD}$에 GAT를 적용할 수 있다. $\overline{AD}$ 와 V에서 노드 i와 j 사이의 어텐션 계수 e는 학습 가능한 가중치 행렬 W_c를 사용하여 다음 수학식 9와 같이 얻을 수 있다.

수학식 9

[0164] $$e_{ij} = \mathrm{FN}(W_c \cdot v_i, W_c \cdot v_j)$$

[0166] 여기서, FN은 입력을 $R^{1\times dh} \times R^{dh\times 1}$ →R로 변환하는 간단한 단층 피드포워드 신경망이다. 위의 표현은 노드 i-j의 특징의 중요성을 나타낼 수 있다. 이때, j는 모든 노드를 나타내지 않고 노드 i의 인접 노드 N_i만을 나타낸다.

마지막으로 소프트맥스 함수를 통과하면 다음과 같은 정규화된 어텐션 행렬 $\widetilde{AD}$를 계산할 수 있다.

수학식 10

[0168] $$\widetilde{AD}_{ij} = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})}$$

[0170] 노드 특징의 다양성을 보장하기 위해 모든 인코더 계층에 위의 수학식 10을 적용할 수 있다. 최종 GHA는 GAT의 노드 특징 행렬 V와 GAT의 가중치 행렬 W_{gat}을 정규화된 어텐션 행렬 $\widetilde{AD}$에 적용하여 다음 수학식 11과 같이 생성될 수 있다.

수학식 11

[0172] $$\mathrm{GHA}(V) = \sigma_2(\widetilde{AD} \cdot V \cdot W_{gat})$$

[0174] 여기서 σ_2는 ReLU 활성화 함수이다. GHA의 값은 도 1의 (c)의 과정을 통해 구할 수 있다. GHA 인코더는 L층으로 구성되어 있기 때문에 제1층에는 V가 입력되지만, 제2층부터는 각층의 출력인 h_l로 변경될 수 있다. 이전 계층의 GHA(h_{l-1})는 다시 입력 h_{l-1}과 스킵 연결될 수 있다(element-wise sum).

수학식 12

[0176] $$\acute{h}_l = \mathrm{GHA}(h_{l-1}) + h_{l-1}, l \in \{1 \cdots L\}$$

[0178] $\acute{h}_l$을 다시 선형 정규화(LN)하여 FFN에 적용한 후, 도 1의 (a)에서와 같이, 원래의 $\acute{h}_l$과 유사하게 스킵 연결하여 최종 인코더 블록의 출력을 생성할 수 있다.

수학식 13

[0180] $$h_l = \mathrm{FFN}(\acute{h}_l) + \acute{h}_l, l \in \{1 \cdots L\}$$

[0182] 마지막으로, 마지막 인코더 계층 L의 출력인 $h_L \in \mathbb{R}^{P \times d_h}$가 판독 계층(readout layer)으로 전달될 수 있다. 판독 계층(readout layer)에서는, 다양성을 고려하기 위해 시퀀스 풀링(seq)(Hassani et al., 2021), 평균 및 최대값이 사용될 수 있다. 다중-판독 특징 H_{out}은 다음 수학식 14와 같이 계산될 수 있다.

수학식 14

[0184] $$H_{out} = h_{seq} \parallel h_{mean} \parallel h_{max}$$

[0186] 여기서 ∥는 노드 특징 간의 연결 작업을 나타낸다. 판독 계층을 통과한 다중 판독 특징 H_{out}은 MLP를 사용하여 분류될 수 있다. 손실 함수는 soft distillation을 사용하여 최적화될 수 있다.

[0188] 도 4는 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법의 흐름을 도시한 도면이다. 도 4에 도시된 바와 같이, 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치의 동작 방법은, GHA-ViT를 사전 학습 없이 소규모의 데이터셋으로 학습해 분류 모델을 생성하는 단계(S100) 및 입력 이미지를 분류 모델에 입력해 분류하는 단계(S200)를 포함하여 구현될 수 있다.

[0190] 각각의 단계들과 관련된 상세한 내용들은, 앞서 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치와 관련하여 충분히 설명되었으므로, 상세한 설명은 생략하기로 한다.

[0192] **실험**

[0194] 제안된 GHA-ViT 모델의 표현 학습 능력을 평가하기 위해 대표적인 CNN 모델인 ResNet(He et al., 2016) 및 MobileNetV2(Sandler et al., 2018), ViT 기반 방법(Dosovitsky et al., 2021; Touvron et al., 2021; Hassani et al., 2021; Yu et al., 2022; MLTin et al021; Touvron et al., 2021a; Liu et al., 2021a); 및 그래프 기반 방법(Han et al., 2022)을 비교하였다. 실험을 통해 제안된 모델의 성능이 여러 벤치마크 데이터셋에서 다른 최첨단(SoTA) 방법의 성능과 유사함을 증명한다. 또한 절제 실험을 통해 그래프 생성 모듈의 효과를 입증하고 어텐션 시각화를 통해 더 적은 헤드를 사용하더라도 유망한 성능을 달성할 수 있음을 입증했다.

[0196] 실험 셋업

[0197] 모델의 용량을 측정하기 위해 다양한 벤치마크 분류 데이터셋을 사용하였다. CIFAR10/100(Krizhevsky et al., 2009), MNIST 및 MNIST-Fashion은 소규모 데이터셋으로 사용하고, ImageNet-1K(Deng et al., 2009)는 중규모 데이터셋으로 사용하였다.

[0199] 또한, 제안된 GHA-ViT 모델이 헤드 및 인코더 계층의 수를 줄일 수 있음을 증명하기 위해 GHA의 기준선(baseline)을 다르게 설정하였다. GHA-ViT의 기본 구조는 본질적으로 소규모 데이터셋으로 학습할 수 있기 때문에 DeiT(Touvron et al., 2021b)를 기반으로 한다. 기준 모델은 헤드와 레이어의 수에 따라 GHA-Base(GHA-B)와 GHA-Small(GHA-S)의 두 가지 유형으로 구성하였다.

[0201] 도 5는 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치에서, 실험에 사용된 기준 GHA-ViT 모델의 변형을 요약하여 도시한 도면이다. 여기서, dim d는 인코더의 잠재 차원(hidden dimension)을 의미하며, mlp ratio는 MLP의 잠재 차원에 대한 스케일링 계수(scaling factor)를 의미한다. GHA-*-a/b에서 a는 계층(layer) 수, b는 패치 크기를 각각 나타낸다.

[0203] GHA-ViT 외에도 비교 실험을 위한 CNN 기준 모델로 ResNet(He et al., 2016)을 사용했다. ResNet 모델은 각 실험 데이터셋의 클래스 수에 맞게 마지막 MLP 계층을 수정했다. 순수한 ViT와 DeiT는 GHA-ViT와 비교하기 위해 트랜스포머 모델로 사용된다. 이러한 방법은 또한 각 실험 데이터셋에 적합한 출력을 얻기 위해 마지막 MLP 레이어를 변경했다.

[0205] 실험환경 및 파라미터 설정

[0206] Pytorch 프레임워크를 사용하여 GHA-ViT 모델을 구축했다. 학습에는 NVIDIA 3090Ti가 사용되었다. 다음 매개변수가 모델 교육에 사용되었다. Optimizer는 AdamW를 사용하였으며, 값은 $\beta 1 = 0.9$, $\beta 1 = 0.999$로 설정하였다. 배치 크기는 모든 학습 및 테스트에 대해 128로 설정되었다. 초기 학습 속도는 0.005로 설정하였으며, 코사인 warm-up과 decay를 사용하였다. 학습을 위해 에포크는 300으로 설정되었고 ImageNet의 입력 해상도는 224×224였으며, 소규모 데이터셋에는 32×32 해상도가 사용되었다.

[0208] 최첨단 모델과 성능 비교

[0209] 도 6은 소규모 데이터셋에서 스크래치 훈련된 CNN과 트랜스포머 기반 모델의 성능을 비교하여 나타낸 도면이다. 여기서, 이미지 해상도는 32×32로 동일하다. †는 순수한 ViT로서, 12개의 멀티헤드와 16개의 커널 크기로 구성되어 있다. 그러나 이미지 크기가 32×32이기 때문에 커널의 크기는 다른 방법과 같은 방식으로 4로 감소해 실험하였다. ††는 ViT-Lite 버전으로, 32×32 크기의 입력 이미지에 서로 다른 패치 크기를 사용하였다.

[0211] GHA-ViT(GHA-B 및 GHA-S)는 모든 유형의 소규모 데이터셋에 대해 SoTA 방법을 능가했다. 또한 GHA-ViT는 학습 매개변수와 운영 측면에서 효율적으로 작동하는 것으로 나타났다. 제안된 GHA-ViT의 작은 버전인 GHA-S는 CIFAR10과 100에서 SoTA 방법을 능가했다. GHA-B는 MNIST에서 ResNet-18보다 0.1%, MNIST-F에서 Vit12/4보다 0.1% 낮은 값을 각각 보여주었지만 CIFAR10과 CIFAR100 데이터셋에서는 최고의 성능을 달성했다. 트랜스포머 기반 모델(Dosovitsky et al., 2021; Hassani et al., 2021)은 낮은 유도 편향으로 인해 소규모의 데이터셋을 사용한 환경에서 전반적으로 낮은 성능을 보였다. 그러나 본 발명의 GHA-ViT 모델은 제안된 그래프 구조를 사용할 때 이미지의 위치와 전역성을 유지함으로써 작은 크기의 데이터셋에서 우수한 성능을 보였다. 매개변수 및 연산 수 측면에서 GHA-S는 순수 ViT-12/4보다 약 17배 적은 매개변수를 사용하고 약 5.8배 적은 연산을 수행하지만 CIFAR100 데이터셋에서 4.3% 높은 정확도를 보여준다. 결과를 통해, 본 발명의 GHA-ViT 모델은 기존 CNN 방법만큼 높은 유도 편향을 보장하고, 적은 수의 훈련 데이터로도 효율적으로 작동한다는 것을 알 수 있다.

[0213] 또한, 제안된 모델이 소규모 데이터셋뿐만 아니라 중간 크기 데이터셋에서도 효과적으로 작동한다는 것을 입증하기 위해 ImageNet-1K 데이터셋을 사용한 실험을 수행했다.

[0215] 도 7은 ImageNet-1K 데이터셋에 대한 본 발명의 GHA-ViT 방법과 SoTA 방법의 성능 비교 결과를 나타낸 도면이다. CNN 기반 방법 중 ResNet-152(He et al., 2016)와 비교하면 GHA-S 모델의 매개변수 수는 최대 6배 감소하고 연산 수는 최대 6.4배 빠르다. 정확도 측면에서 GHA-S 모델은 ResNet-152보다 약간 떨어지는 반면, GHA-B 모델은 Top-1 정확도에서 3.4%의 정확도를 향상시켰다. ViT 기반 방법(Dosovitskey et al., 2021; Touvron et al., 2021b; Hassani et al., 2021; Yuan et al., 2022)과 비교하여, GHA-S 모델은 동일한 조건(Param 및 MACs)에서 Top-1 정확도를 0.5% 증가시켰다. GHA-B의 경우 Top-1 정확도가 81.7%로 가장 높게 나타났으나, 유사한 정확도에서 T2T-ViT-14에 비해 1.1배 많은 연산 수를 보였다. MLP 기반 모델(Tolstikhin et al., 2021; Touvron et al., 2021a; Liu et al., 2021a)과 비교하여, GHA-S와 GHA-B 모델 모두 매개변수와 연산 수가 유사한 gMLP-Ti와 gMLP-S에서 Top-1 정확도를 약 2% 증가시켰다. 또한 그래프 구조를 이용한 ViG 방법(Han et al., 2022)과 비교하면, GHA-S와 GHA-B 모델은 매개변수와 연산 수가 약간 많은 반면 각각 3.8%, 1.3% 더 높은 정확도를 보였다. 제안된 GHA-ViT 모델은 ViG에서 사용되는 그래프 생성 방법보다 높은 효율의 그래프 구조를 생성할 수 있기 때문이다.

[0217] 애플레이션 연구 1 - 그래프 구조 생성

[0218] 도 8은 그래프 구조 생성 방법에 따른 정확도(%)를 비교하여 나타낸 도면이다. 본 실험에서는, 제안된 그래프 생성 방법의 효율성을 검증하기 위해 도 8과 같이 다양한 데이터셋의 그래프 구조 변화에 따른 정확도의 차이를 관찰하였다. 여기서 $\mathcal{G}_{w/o}$는 어텐션 맵에서 Top-k 방법만을 사용하여 그래프를 만들고, $\mathcal{G}_{tirU}$는 상삼각 행렬을 사용하여 그래프를 만들며, $\mathcal{G}_{tirL}$은 하삼각 행렬을 사용하여 그래프를 만든다. 상삼각 행렬 및 하삼각행렬을 모두 사용해 그래프를 생성하는 $\mathcal{G}_{all}$의 경우, 모든 데이터셋에서 다른 세 가지 경우보다 더 높은 정확도를 얻었다. 실험을 통해 어텐션 맵의 특정 영역만을 고려하여 그래프 구조를 만들면 그래프가 이미지의 전체적인 특성을 정확하게 반영할 수 없음을 알 수 있다.

[0220] 애플레이션 연구 2 - 다중-판독(multi-readout) 특징의 효과

[0221] 도 9는 다중-판독 특징의 효과를 입증하기 위해 소규모 데이터셋에서 서로 다른 판독 계층의 정확도(%)를 비교해 나타낸 도면이다. 도 9는 제안된 다중-판독 특징이 모든 소규모 데이터셋에서 가장 높은 성능을 나타냄을 나타낸다. 결과적으로, 결합된 다중-판독 특징이 최대 또는 평균 연산과 같은 간단한 연산을 통해 얻은 판독 특징만 사용하는 것보다 최종 이미지를 더 효과적으로 분류하는 데 도움이 된다는 것을 알 수 있다.

[0223] 애플레이션 연구 3 - 그래프 컨볼루션의 유형

[0224] 도 10은 소규모 데이터셋에 대한 GHA-S에 여러 유형의 그래프 컨볼루션을 적용하였을 때의 정확도(%)를 비교하여 나타낸 도면이다. 본 실험에서는, 대표적인 그래프 컨볼루션 방법으로 GCN(Kipf & Welling, 2016), GIN(Xu et al., 2018), GAT(Veli$\check{c}$kovi$\acute{c}$ et al.)를 사용하였다. 도 10에서 보는 바와 같이, 제안된 GHA-ViT 방법은 다양한 그래프 컨볼루션 연산에서 큰 차이를 보이지 않는다. 이것은 GHA가 본질적으로 매우 유연한 구조를 가지고 있다는 것을 의미한다. 3개의 그래프 컨볼루션 중 GAT가 패치 간의 모든 관계를 동일한 방식으로 처리하는 GCN이나 GIN과 달리 패치 간의 관계를 다르게 처리하기 때문에 GAT를 사용했을 때 가장 높은 정확도를 얻었다. 본 연구에서는 실험 결과를 고려하여 기본 그래프 컨볼루션으로 GAT를 사용하였다.

[0226] 그래프 시각화

[0227] 도 11은 본 발명의 일실시예에 따른 소규모의 데이터셋을 위한 그래프 헤드 어텐션(100) 기반의 비전 트랜스포머 장치 및 이의 동작 방법에서 생성된 그래프 구조를 시각화하여 나타낸 도면이다. 여기서, (a)는 입력 이미지 패치에 k-NN을 직접 적용하여 인접 행렬을 생성하는 ViG(Han et al., 2022)의 그래프 생성 방법에 따른 인접 행렬, (b)는 GHA-S에 의해 생성된 인접 행렬을 각각 나타내며, 빨간색 선은 이미지의 타겟 패치(빨간색 별 및 상자 영역)에 해당하는 행을 나타낸다. 인접 행렬의 오른쪽에 있는 패치는 타겟 패치(노란색 삼각형 및 상자 영역)와 가장 높은 관계가 있는 패치이다. 도 11에서 보는 바와 같이 k-NN을 통해 생성된 그래프 구조는 타겟 패치와 관련이 없는 배경 패치를 포함하는 반면에, 본 발명에서 제안하는 방법은 타겟 패치와 관련성이 높은 패치만 선택하는 것을 확인할 수 있다. 이는 이미지 기반 k-NN 방식보다 어텐션 맵을 이용한 그래프 생성 방식이 더 효과적임을 나타낸다.

[0229] 본 발명에서는 ViT의 핵심 모듈인 MHA의 한계를 극복할 수 있는 새로운 GHA 방법을 제안하였다. 어텐션 맵 연산을 행렬 관점에서 그래프 관점으로 변환함으로써 영상분류의 정확도를 유지하면서 불필요한 연산과 파라미터의 수를 크게 줄일 수 있었다. 또한, 멀티-헤드에 내장된 어텐션 특징 공간이 그래프 헤드를 더 적게 사용했을 때와 크게 다르지 않다는 것을 증명했다. 향후 추가 연구를 통해 GHA 구조의 ViT 성능이 MHA 구조의 ViT 성능보다 크게 향상될 것으로 기대된다.

[0231] 본 발명의 실시예는 컴퓨터에 의해 실행되는 매체에 저장된 컴퓨터 프로그램 또는 컴퓨터에 의해 실행가능한 명

령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비 분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비 분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 메커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.

[0233] 본 발명은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0235] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명된 각 구성요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명된 구성 요소들도 결합된 형태로 실시될 수 있다.

[0237] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0238] 100: 그래프 헤드 어텐션

110: 그래프 생성부

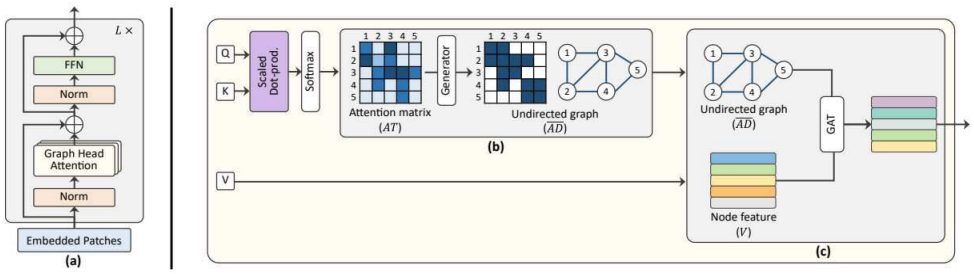
120: 어텐션 생성부

S100: GHA-ViT를 사전 학습 없이 소규모의 데이터셋으로 학습해 분류 모델을 생성하는 단계

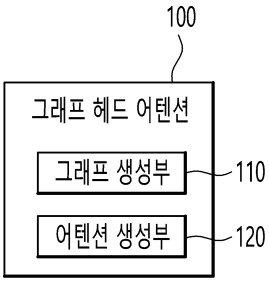
S200: 입력 이미지를 분류 모델에 입력해 분류하는 단계

도면

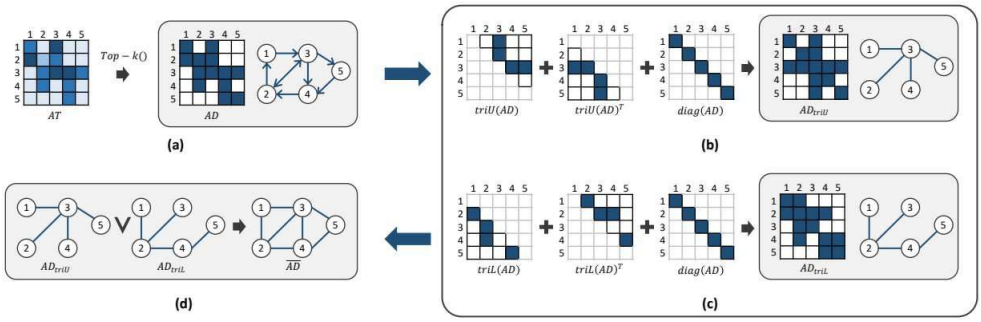
도면1



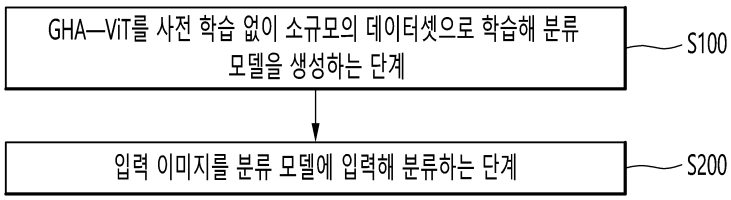
도면2



도면3



도면4



도면5

Model	head	layers	dim d	mlp ratio
GHA-S-7/3	4	7	64	2
GHA-B-7/3	6	7	64	2
GHA-S-14/7	3	14	64	4
GHA-B-14/7	6	14	64	4

도면6

Model	Params (M)	MACs (M)	Top-1(%)			
			CIFAR10	CIFAR100	MNIST	MNIST-F
ResNet-18 (He et al., 2016)	11	40	90.2	66.4	99.8	94.7
ResNet-34 (He et al., 2016)	21	80	90.5	66.8	99.7	94.7
MobileNetV2/0.5 (Sandler et al., 2018)	1	10	84.7	56.3	99.7	93.9
MobileNetV2/2.0 (Sandler et al., 2018)	8	20	91.0	67.4	99.7	95.2
ViT-12/4† (Dosovitskiy et al., 2021)	85	5520	94.8	74.1	99.6	95.4
ViT-Lite-7/16†† (Hassani et al., 2021)	3	20	78.4	52.8	99.6	93.2
ViT-Lite-7/8†† (Hassani et al., 2021)	3	60	89.1	67.2	99.6	94.4
ViT-Lite-7/4†† (Hassani et al., 2021)	3	260	93.5	73.9	99.7	95.1
CVT-7/8 (Hassani et al., 2021)	3	60	89.7	70.1	99.7	94.5
CVT-7/4 (Hassani et al., 2021)	3	250	94.0	76.4	99.7	95.3
CVT-7/3×2 (Hassani et al., 2021)	3	1290	95.0	77.7	99.7	95.1
GHA-S-7/3	5	950	95.2	78.4	99.5	95.2
GHA-B-7/3	10	2130	96.8	80.1	99.7	95.3

도면7

Model	Params (M)	MACs (G)	Top-1 (%)	Top-5(%)
ResNet-50 (He et al., 2016)	26	4.3	76.2	95.0
ResNet-101 (He et al., 2016)	45	7.9	77.4	95.4
ResNet-152 (He et al., 2016)	60	11.6	78.3	95.9
ViT-S-16 (Dosovitskiy et al., 2021)	47	10.1	78.1	-
DeiT-S (Touvron et al., 2021b)	22	4.6	79.8	95.0
CCT-14/7×2 (Hassani et al., 2021)	22	18.6	80.6	-
T2T-ViT-14 (Yuan et al., 2021b)	22	4.8	81.5	-
PoolFormer-S12 (Yu et al., 2022)	12	1.8	77.2	-
PoolFormer-S24 (Yu et al., 2022)	30	3.0	80.3	-
Mixer-B/16 (Tolstikhin et al., 2021)	59	12.7	76.4	-
ResMLP-12 (Touvron et al., 2021a)	15	3.0	76.6	-
gMLP-Ti (Liu et al., 2021a)	6	1.4	72.3	-
gMLP-S (Liu et al., 2021a)	20	4.5	79.6	-
ViG-Ti (Han et al., 2022)	7	1.3	73.9	92.0
ViG-S (Han et al., 2022)	23	4.5	80.4	95.2
GHA-S-14/7	10	1.8	77.4	93.5
GHA-B-14/7	29	5.9	81.7	95.8

도면8

Dataset	$\mathcal{G}_{w/o}$	$\mathcal{G}_{triU}$	$\mathcal{G}_{triL}$	$\mathcal{G}_{all}$
CIFAR10	93.2	93.1	92.8	95.2
CIFAR100	73.8	73.7	73.0	78.4
MNIST	99.4	99.3	99.4	99.5
MNIST-F	94.8	94.7	94.8	95.2

도면9

Dataset	mean	max	seq	multi
CIFAR10	93.9	92.5	91.1	95.2
CIFAR100	73.8	72.1	75.1	78.4
MNIST	99.2	99.4	99.2	99.5
MNIST-F	94.8	93.4	94.9	95.2

도면10

Dataset	GCN	GIN	GAT
CIFAR10	94.5	94.9	95.2
CIFAR100	76.7	77.1	78.4
MNIST	99.4	99.4	99.5
MNIST-F	94.8	94.9	95.2

도면11

