최신 비전 트랜스포머 변형 모델 동향에 관한 연구

지훈*, 임한비**, 이연준***
*한양대학교 컴퓨터공학과(대학원생), **한양대학교 정보보호학과(대학원생), ***한양대학교 컴퓨터공학과(교수)

*greenpea0819@hanyang.ac.kr, **hanbi1@hanyang.ac.kr, ***yeonjoonlee@hanyang.ac.kr

A Survey on Recent Trends in Vision Transformer Variants

Hoon Ji*, Hanbi Yim**, Yeonjoon Lee***
*Hanyang University, Computer Science and Engineering (Master Student)
**Hanyang University, Department of Computer and Information Security (Master Student)
***Hanyang University, Computer Science and Engineering (Associate Professor)

요 약

Vision Transformer (ViT)는 기존의 합성곱 신경망(CNN) 중심의 패러다임에 새로운 방향을 제시하며 컴퓨터 비전 분야에 혁신을 가져왔다. 그러나, 입력 크기에 따른 2 차 연산 복잡도와 대규모 데이터셋에 대한 의존성이라는 명확한 한계를 지니고 있기에, ViT 의 아키텍처를 변형하여 이를 극복하려는 연구들이 활발히 진행되고 있다. 본 논문은 이러한 변형 ViT 모델들을 아키텍처 설계 전략에 따라 분류하고, 이에 속한 연구들을 분석한다. 본 논문에서는 변형 ViT 모델들을 접근법에 따라, (1) 어텐션 구조 효율화, (2) CNN-Tranformer 하이브리드 아키텍처, 그리고 (3) 토큰 처리 효율화라는 세 가지 기준으로 분류하고 최신 연구 동향을 조망한다. 각 분류에 속하는 주요 모델들의 핵심 아이디어와 기술적 차별점에 대해 분석함으로써 ViT 분야의 주요 해결 과제와 발전 방향에 대한 통찰을 제공하는 것을 목표로 한다. 나아가, ViT 가 목표로 삼아야 하는 주요 과제에 대해 제시함으로써 추후 연구에 관한 방향성을 제시한다.

Ⅰ. 서 론

Vision Transformer (ViT)는 자연어 처리에 특화된 트랜스포머 구조를 비전 태스크에 적용하여 [1], 기존의 합성곱 신경망 (CNN) 중심의 패러다임에 새로운 방향을 제시하였다. ViT 는 셀프 어텐션 매커니즘을 통해 이미지의 전역적인 문맥을 효과적으로 학습할 수 있음을 보여주었으나, 입력 크기에 따른 2 차 연산의 복잡도와 대규모 데이터셋 의존성이라는 명확한 한계점을 지니고 있으며, 이를 해결하기 위한 후속 연구들이 활발히 수행되고 있다 [2]. 따라서, 수많은 후속 연구들을 적절히 분류하고 분석할 수 있는 기준을 마련하는 것은 필수적이다.

본 논문은 최신 ViT 변형 모델들의 동향을 조망하고, 이들의 핵심적인 아키텍처 설계 전략에 대해 분석한다. 이를 위해, 우리는 ViT 의 한계점을 해결하는 전략에 따라 후속 연구들을 크게 (1) 어텐션 구조 효율화, (2) CNN-Transformer 하이브리드 아키텍처, 그리고 (3) 토큰 처리 효율화로 분류한다. 우리는 각 분류에 속하는 주요 모델들의 핵심 아이디어와 기술적 차별점에 대해 분석함으로써, ViT 분야의 동향 및 후속 연구의 방향에 대한 통찰을 제공하는 것을 목표로 한다.

Ⅱ. 본론

본론에서는 앞서 제시한 세 가지 분류 기준에 따라 최신 변형 ViT 관련 연구들을 분류하고, 각 연구들의 핵심적인 아이디어와 접근 방식을 살펴본다.

1. 어텐션 구조 효율화

초기 ViT 의 셀프 어텐션은 모든 토큰 간의 관계를 계산하기에, 입력 이미지의 크기가 커짐에 따라 연산량이 기하급수적으로 증가하는 2 차 복잡도 문제를 내재하고 있다. 이를 해결하기 위해, 어텐션의 계산 범위를 전체가 아닌 특정 지역으로 한정하거나 계층적 구조를 도입하여 연산 효율을 극대화하려는 연구들이 수행되었다

Liu et al. [3]은 이미지를 겹치지 않는 윈도우로 분할하고 그 내부에서만 어텐션을 수행하며, 연속된 블록 간에는 윈도우를 이동시켜 정보의 연결성을 확보하였다. 이러한 계층적 설계는 연산량을 선형 복잡도로 감소시켰을 뿐만 아니라, 다양한 비전 태스크의 핵심 백본으로 자리매김하였다. 이와 유사하게, Chu et al. [4]은 각 블록을 지역 그룹 어텐션과 다운샘플링된 토큰을 활용한 전역 어텐션으로 순차 구성했으며, Zhanget al. [5]은 NLP 의 Longformer 구조를 차용하여 밀집된 지역 어텐션과 희소한 전역 어텐션을 결합했다. 이러한 지역-전역 어텐션 분리 전략은 모델이 세밀한 지역 패턴과 넓은 문맥을 동시에 효율적으로 포착할 수 있게 한다. 한편 몇몇 연구들은 다중 해상도 정보를 병렬적으로 처리하는 구조를 제시했다. Yuan et al. [6]은 모델 전체에 걸쳐 여러 해상도의 스트림을 유지하며 지속적으로 정보를 교환해 고해상도 특징을 보존했으며, Lee et al. [7]은 각 단계에서 여러 스케일의 특징을 동시에 추출하는 다중 경로 구조를 통해 더욱 풍부한 공간 정보를 학습하도록 설계했다.

2. CNN-Transformer 하이브리드 아키텍처

기존 ViT 는 지역적 귀납 편향이 부족하여 데이터 의존성이 높다는 한계를 가지므로, 풍부한 귀납 편향을

통해 데이터 효율성을 확보한 CNN 의 장점을 결합하는 하이브리드 아키텍처가 활발히 연구되었다. 이러한 접근법은 CNN 으로 추출한 특징 맵을 ViT 의 입력으로 사용하는 직렬 방식과, 두 아키텍처에서 독립적으로 추출한 특징을 융합하는 병렬 방식으로 크게 나뉜다.

Guo et al. [8]은 모델 초반부에 CNN 으로 지역적인 특징을 추출한 후, 후반부에서 Transformer 를 통해 전역적인 관계를 학습하는 직렬 융합 모델을 제시한다. 반면, 병렬 융합 방식의 연구들인 Xu et al. [9]은 토큰화 및 다운샘플링 과정에 컨볼루션 연산을 통합하여 자연스러운 계층 구조를 형성했으며, Xia et al. [10]은 트랜스포머 블록 내에 다중 스케일 컨볼루션 분기를 병렬로 연결하여 지역 및 전역 정보의 상호작용을 극대화하는 데 집중했다. 추가적으로, Fan et al. [11]은 별도의 컨볼루션 모듈과 어텐션 모듈이 양방향으로 정보를 교환하는 방식으로 두 방식의 장점을 극대화하는 구조를 제시하였다.

3. 토큰 처리 효율화

앞서 말한 두 방식과 다르게, 모델의 연산량을 근본적으로 줄이기 위해 이미지의 내용에 따라 입력 토큰의 수를 동적으로 조절하는 연구들 또한 등장하였다. Yin et al. [12]은 어텐션 스코어를 기반으로 중요도가 낮은 토큰을 다음 레이어의 연산에서 제외시키는 '토큰 가지치기' 방식을 제안하였다. 이 접근법은 이미지의 배경과 같이 정보량이 적은 영역의 계산을 효율적으로 생략한다. 반면, Norouzi et al. [13]은 유사한 토큰들을 점진적으로 병합하는 구조를 제시한다. 초기 레이어에서는 인접 토큰을 합쳐 지역적 특징을 응축하고, 후기 레이어에서는 의미적으로 유사한 토큰들을 찾아 병합함으로써 정보 손실을 최소화하며 토큰 수를 줄이는 방식을 사용한다.

이처럼 입력에 따라 계산량을 조절하는 동적 처리 방식은 ViT 의 연산 패러다임을 고정적 구조에서 적응형 구조로 전환하여 모델의 실용성을 크게 높였다.

Ⅲ. 결론

본 논문은 초기 ViT 가 지닌 2 차 연산 복잡도와 데이터 의존성 문제를 해결하기 위한 ViT 변형 모델에 관한 최신 연구 동향을 세 가지 분류를 통해 분석하였다. 어텐션 구조 효율화는 연산 범위를 제한하여 계산 효율을 높이는 방식을 제안하였고, CNN-Transformer 하이브리드 아키텍처는 CNN 의 귀납 편향을 Transformer 에 결합하여 데이터의 효율성을 개선하였다. 마지막으로 토큰 처리 효율화는 입력에 따라 동적으로 계산량을 조절하는 새로운 패러다임을 제시하였다. 이러한 연구 흐름들은 ViT 가 단순 이미지 분류를 넘어 비전 태스크의 핵심 아키텍처로 발전하는 데 기여하였다. 향후 연구는 이러한 개별적인 효율화 기법들을 유기적으로 통합하여 시너지를 극대화하고, 제한된 자원을 가진 엣지 디바이스에서도 효과적으로 구동될 수 있는 초경량 모델을 개발하는 방향으로 나아가야 할 것이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원(IITP-2025-RS-2024-00438056, 50%)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(RS-2024-00341722, 지능형 서비스 로봇의 사이버 레질리언스 확보를 위한 보안기술 개발, 50%)

참 고 문 헌

[1] Alexey Dosovitskiy et al., "An Image is Worth 16x16Words: Transformers for Image Recognition at Scale",arXiv preprint arXiv:2010.11929, 2020

[2] K. Han et al., "A Survey on Vision Transformer,"IEEETransactions on Pattern Analysis and MachineIntelligence, vol. 45, no. 1, pp. 87-108, 2023

[3] Ze Liu, et al. "Swin transformer: Hierarchical visiontransformer using shifted windows."Proceedings of theIEEE/CVF international conference on computer vision,2021

[4] Xiangxiang Chu et al. "Twins: Revisiting the design ofspatial attention in vision transformers."Advances inneural information processing systems 3, 9355-9366,2021

[5] Pengchuan Zhang et al. "Multi-scale vision longformer:A new vision transformer for high-resolution imageencoding."Proceedings of the IEEE/CVF internationalconference on computer vision, 2021

[6] Yuhui Yuan et al. "Hrformer: High-resolution visiontransformer for dense predict."Advances in neuralinformation processing systems 34, 7281-7293, 2021

[7] Youngwan Lee et al. "Mpvit: Multi-path visiontransformer for dense prediction."Proceedings of theIEEE/CVF conference on computer vision and patternrecognition, 2022

[8] Jianyuan Guo et al. "Cmt: Convolutional neural networksmeet vision transformers."Proceedings of the IEEE/CVFconference on computer vision and pattern recognition,2022

[9] Yufei Xu et al. "Vitae: Vision transformer advanced byexploring intrinsic inductive bias."Advances in neuralinformation processing systems 34, 28522-28535, 2021

[10] Chunlong Xia et al. "Vit-comer: Vision transformerwith convolutional multi-scale feature interaction fordense predictions."Proceedings of the IEEE/CVFconference on computer vision and pattern recognition,2024

[11] Qihang Fan et al. "Lightweight vision transformer withbidirectional interaction."Advances in Neural InformationProcessing Systems 36, 15234-15251, 2023

[12] Hongxu Yin et al. "A-vit: Adaptive tokens for efficientvision transformer."Proceedings of the IEEE/CVFconference on computer vision and pattern recognition,2022

[13] Narges Norouzi et al. "Algm: Adaptive local-then-global token merging for efficient semantic segmentationwith plain vision transformers."Proceedings of theIEEE/CVF Conference on Computer Vision and PatternRecognition, 2024