특집논문 (Special Paper)
방송공학회논문지 제29권 제3호, 2024년 5월 (JBE Vol.29, No.3, May 2024)
https://doi.org/10.5909/JBE.2024.29.3.291
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

적응적 채널 분할 방법을 통한 저복잡도 비전 트랜스포머

이 세 영[a)], 오 상 수[a)], 한 상 현[a)], 임 승 환[a)], 이 종 석[a)], 심 동 규[a)‡]

Low-Complexity Vision Transformer with Adaptive Channel Partitioning Method

Seyoung Lee[a)], Sangsoo Oh[a)], Sanghyun Han[a)], Seunghwan Lim[a)], Jongseok Lee[a)], and Donggyu Sim[a)‡]

요 약

최근 컴퓨터 비전 연구에서는 비전 트랜스포머가 백본 네트워크로써 높은 성능을 입증했지만, 높은 네트워크 복잡도로 인해 에지 장치로의 배포는 여전히 제한적이다. 본 논문에서는 에지 장치로의 적용을 용이하게 하기 위해, 적응적 채널 분할 방법을 통한 저복잡도 비전 트랜스포머 모델을 제안한다. 본 논문에서는 분할할 채널의 비율과 레이어의 민감도에 따라 제안 방법을 적응적으로 적용하여, 모델의 연산량 및 메모리 부담을 개선하면서도 정확도 손실율을 최소화한다. 본 논문의 5-겹 교차 검증의 실험 결과, 제안하는 저복잡도 비전 트랜스포머는 기존 비전 트랜스포머 모델 대비 FLOPs 약 24.9%, 파라미터 수 약 40.4%, 추론 속도 약 16.7%의 개선을 보였으며, 분류 정확도는 Top 1 Accuracy 기준 최대 1.18%p 향상되었다.

Abstract

Recent research in computer vision has demonstrated the high performance of Vision Transformers as backbone networks; however, their deployment to edge devices remains limited due to their high network complexity. In this paper, we propose a low-complexity Vision Transformer model facilitated for edge device deployment through an adaptive channel partitioning method. The proposed method adaptively applies according to the sensitivity of the layers and the ratio of channels partitioned, thereby improving the model's computational and memory burden while minimizing accuracy loss. Our experimental results, using 5-fold cross-validation, showed a reduction of about 24.9% in FLOPs, 40.4% in the number of parameters, and a 16.7% improvement in inference speed compared to existing Vision Transformer models, with an increase in classification accuracy of up to 1.18 percentage points in terms of Top 1 Accuracy.

Keyword: Deep learning, Vision transformer, Low complexity, Adaptive channel partitioning, Edge device

Ⅰ. 서 론

최근 기계 번역, 문서 요약, 챗봇과 같은 자연어 처리 (NLP) 분야에서 우수한 대용량 병렬 처리 및 학습 능력을 보인 트랜스포머 (Transformer) 구조가 딥러닝 기반의 컴퓨터 비전 작업에서 합성곱 신경망 (Convolutional Neural Network, CNN)과 더불어 주류 모델로 자리잡았다[1-3]. 컴퓨터 비전 분야로 확장 적용된 비전 트랜스포머 (Vision Transformer, ViT) 부류의 모델은 어텐션 (Attention) 메커니즘을 통해 컨볼루션 연산의 국소적 수용 영역의 한계를 극복하고 이미지 전체 영역에 대한 광역적 컨텍스트 정보를 효과적으로 학습할 수 있다[2][4-9]. 실제로 비전 트랜스포머 모델들은 대용량 학습 데이터를 기반으로 이미지 분류 (Image classification)와 객체 검출 (Object detection), 객체 추적 (Object Tracking), 영상 분할 (Image segmentation) 등의 컴퓨터 비전 벤치마크에서 SOTA (State-of-the-art) 성능을 달성하였다[10-13]. 또한, 영상의 시각 품질을 개선하기 위한 디노이징 (Denoising), 초해상화 (Super-resolution), 프레임율 상향 변환 (Frame rate up conversion) 기술들은 비전 트랜스포머를 접목하여 높은 정확도를 갖추게 되었으며, 비전 트랜스포머를 사용한 영상 압축 기술은 VVC (Versatile Video Coding, H.266) 대비 더 낮은 비트율-왜곡 비용을 보여주었다[14-18]. 그러나 트랜스포머 기반의 모델은 특유의 높은 계산 복잡도로 인해 많은 연산량과 메모리를 요구하기 때문에, 합성곱 신경망에 비해 에지 장치에 통합하기에 어려움이 있다[19-21]. 특히 트랜스포머의 핵심 구성 요소인 어텐션 블록은 연산량 관점에서 병목 구간으로 작용하며, 모델의 전체 메모리 사용량 중 가장 큰 비중을 차지한다[21-22]. 이로 인해 트랜스포머 부류 모델의 원활한 동작을 위해서는 고성능 GPU 시스템과 같은 고수준의 컴퓨팅 자원이 요구되며, 결과적으로 저장 공간, 실행 메모리 등 컴퓨팅 자원의 제약이 따르는 에지 장치로의 배포가 제한된다[23-27]. 따라서 모바일 기기, 로봇, 사물 인터넷 (IoT) 등 다양한 에지 장치에서 비전 트랜스포머 기반 모델을 효과적으로 활용하기 위해서는 트랜스포머 모델의 경량화 방법을 모색하고 연산 효율성을 높이는 것이 필수적이다.

딥러닝 모델에 대한 경량화 및 고속화 연구는 모델의 연산 복잡도를 낮추고, 추론 속도를 향상시키며, 메모리 사용량을 감소시켜 임베디드 시스템과 같은 자원이 제한된 환경에서 활용 가능성이 높은 연구 분야이다[28-29]. 기존 연구들은 경량화 목적을 달성하기 위해 가지치기 (Pruning), 양자화 (Quantization), 지식 증류 (Knowledge Distillation) 기법 및 변형된 구조들을 제안하였다[28][30-32]. 첫째로, 가지치기는 불필요한 가중치나 연산을 제거함으로써, 모델의 크기를 줄이고 계산 효율성을 높이며, 일부 경우에는 과적합을 방지할 수 있는 방법이다[33-36]. 예시로, Vision Transformer Pruning (VTP)은 트랜스포머 모델의 멀티 헤드 어텐션 (Multi-head Attention, MHA) 및 다층 퍼셉트론 (Multi-Layer Perceptron, MLP) 블록에서의 부동 소수점 연산 횟수 (FLOPs)를 줄이기 위한 구조적 가지치기 방법을 제안하였다[34]. 둘째로, 양자화는 모델 가중치와 활성화를 낮은 비트 폭으로 표현함으로써, 모델의 크기를 줄이고 추론 속도를 가속화하기 위한 기술이다[37-40]. 일반적으로 32 비트의 부동소수점 가중치를 N 비트의 정수로 변환하는 과정이며, I-Transformer는 비전 트랜스포머를 위한 정수화 방법인 I-ViT를 통해 실수 사용을 제거함으로써 저장 및 연산에 필요한 부담을 줄였다[40]. 셋째로, 지식 증류는 복잡한 교사 모델로부터 얻을 지식을 보다 간단한 학생 모델에 전달하여, 학생 모델의 추론 성능을 향상시키는 방법이다[41-43]. DeiT는 비전 트랜스포머 모델에 특유의 지식 증류 기법인 토큰 기반 전략을 적용하여, 기존 비전 트랜스포머보다 더 적은 데이터와 컴퓨팅 자원으로 우수한 이미지 분류 성능을 달성하였다[43]. 마지막으로, 합성곱 신경망과 트랜스포머를 함께 활용하여 빠른 실행 속도를 보장하는 모델 경량화 방법이 소개되었다[44-49]. CVT 모델은 기존 임베딩 방법에 합성곱 신경망을 접목하였고, LeViT는 합성곱 신경망을 통해 추출한 특징 맵을 트랜스포머 블록에서 활용하는 방

a) 광운대학교 컴퓨터공학과(Department of Computer Engineering, Kwangwoon University)
‡ Corresponding Author : 심동규(Donggyu Sim)
E-mail: dgsim@kw.ac.kr
Tel: +82-2-940-5470
ORCID: https://orcid.org/0000-0002-2794-9932
※본 연구는 대한민국 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2021R1A2C2092848), 2023년 교내학술연구비 지원 및 2024년도 정부(산업통상자원부)의 재원으로 한국산업기술 진흥원의 지원 (P0017124, 2024년 산업혁신인재성장지원사업)을 받아 수행되었음.
· Manuscript April 16, 2024; Revised May 13, 2024; Accepted May 14, 2024.

법과 다운샘플링 방법을 제안하였다[50-51]. MobileViT는 합성곱 신경망과 트랜스포머 블록을 번갈아 사용하는 네트워크를 제시하였다[52].

다양한 모델 경량화 연구와 그 개발은 트랜스포머 부류 모델의 효율성을 증대시키고 에지 장치에서의 사용 가능성을 확대하고 있지만, 트랜스포머 모델 내부의 근본적인 계산 복잡성 문제를 해결하지는 못한다[21-22]. 트랜스포머의 핵심 메커니즘인 어텐션 연산의 계산 복잡도는 입력 시퀀스 길이에 대해 복잡도를 가지기 때문에 입력 데이터의 크기가 증가할수록 모델 동작에 필요한 연산량이 기하급수적으로 커지고, 학습 및 추론 시간이 급격히 증가한다. 기존 연구에서는 어텐션 연산의 계산 복잡도를 낮추기 위해, 여러 저복잡도 어텐션 기법들을 제안하였다[22][53-58]. 그 중 Performer와 Nystromformer, Linformer 등은 선형 계산 복잡도를 달성하였다[53-55]. 이들은 기존 트랜스포머 모델의 전체 연산량을 줄이거나 네트워크 파라미터 수를 줄이면서도 기존 트랜스포머 모델의 높은 정확도를 유지했다. 그러나 여전히 비전 트랜스포머의 대부분 연산량은 멀티 헤드 어텐션 블록이 차지하며, 요구되는 연산량은 입력 데이터의 크기에 비례한다.

본 논문에서는 입력 데이터 크기에 의존적인 연산량 부담을 경감시키기 위하여 합성곱 신경망의 경량화 기법으로 소개되었던 CSP 네트워크의 형태를 트랜스포머의 멀티 헤드 어텐션 블록에 적용하는 적응적 채널 분할 방법을 제안한다[59]. 제안하는 방법은 CSP 네트워크 형태를 하위 레이어에 위치한 어텐션 블록에만 적응적으로 결합함으로써, 트랜스포머 기반 모델의 어텐션 블록에 필요한 메모리 및 연산량을 획기적으로 줄이고 모델 정확도 저하를 최소화하였다. 또한, 본 논문의 실험에서는 트랜스포머 기반 모델들에 대한 제안하는 방법의 범용성과 확장 적용 가능성을 검증하기 위해, 기존 비전 트랜스포머와 선형 계산 복잡도를 가진 어텐션 모델에 대해 제안하는 방법의 효율성을 평가하였고, 합성곱 신경망을 활용하여 모델 경량화를 달성한 하이브리드 네트워크에 대해서도 제안하는 방법의 효과를 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서 기존 트랜스포머 연구에 대해 간략히 설명하고, 3장에서 기존 연구의 효율성을 개선하기 위한 제안 방법을 소개한다. 그리고 4장에서는 다양한 실험을 통해 제안하는 방법의 효과를 평가하고 검증한다. 마지막으로 5장에서 결론을 맺는다.

II. 기존 연구

트랜스포머의 멀티 헤드 어텐션 블록 구조는 그림 1과 같다. 시퀀스 내 각 토큰 요소가 다른 토큰들로부터 유사 정보를 학습하도록 단일 헤드의 연산은 수식 1과 같이 정의된다.

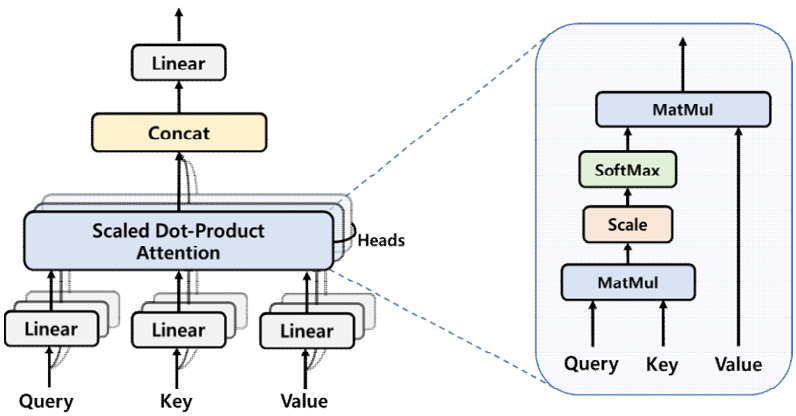


그림 1. 트랜스포머의 멀티 헤드 어텐션과 스케일된 내적 어텐션의 블록 다이어그램
Fig. 1. A block diagram representing a multi-head attention mechanism and scaled dot product attention of Vanilla Transformer

$$Attention_h = Softmax(\frac{1}{\sqrt{d}} Query_h Key_h^T) Value_h \quad (1)$$

이때, N과 d, h는 각각 시퀀스의 길이와 채널 깊이, 헤드 수를 의미하고 $\frac{1}{\sqrt{d}}$은 스케일링 인자이다. 입력 시퀀스 X를 $R^{N\times d}$ 행렬이라 할 때, Linear 블록은 각 헤드의 쿼리 (Query), 키 (Key), 값 (Value)이 $Q_h = XW_q$, $K_h = XW_k$, $V_h = XW_v$로 계산되는 선형 변환이다. W_q와 W_k, W_v는 입력 시퀀스 X를 d차원의 출력 텐서로 투영하기 위한 가중치 행렬을 나타낸다. 어텐션 연산은 수식 1과 같이 쿼리 (Query)와 키 (Key)의 각 토큰 사이의 내적을 취하여 시퀀스 내 토큰들 사이의 유사 관계를 0과 1 사이의 값으로 확률화하고, 다시 값 (Value)과 내적하여 각 토큰들의 어텐션 특징 맵을 표현한다. 결국 정의된 손실 함수에 의해 어텐션 특징 맵이 유의미한 정보를 갖는 텐서가 되도록 각 가중치 행렬인 W_q와 W_k, W_k가 학습된다. 이를 통해 시퀀스 내의 각 토큰이 다른 모든 토큰과의 관계를 파악하게 되므로, 전체 데이터의 광역적 문맥 정보를 효과적으로 학습할 수 있다[4-5].

그러나 행렬 내적 연산 과정에서 $O(N^2)$의 계산 복잡도를 가진 어텐션 메커니즘은 트랜스포머 모델 구조에서 병목 구간으로 작용한다. 이에 따라 어텐션 메커니즘의 시간, 공간 복잡도를 개선하고 컴퓨팅 비용과 메모리 부담을 줄이기 위해 다양한 어텐션 경량화 연구가 진행되었다. 그 중 Performer, Nystromformer, Linformer는 각각 커널화, 다운샘플링, 저 랭크 근사를 활용한 어텐션 모델들로 계산 복잡도를 $O(N)$으로 낮추었다. Performer[53]는 FAVOR+ (Fast Attention Via positive Orthogonal Random feature approach) 메커니즘을 통해 가우시안 커널을 이용하여 소프트맥스 함수를 추정하는 방법을 제안하여 빠른 어텐션 연산을 수행하였다. 그리고 Nystromformer[54]는 가중 행렬의 세그먼트 단위 평균을 통해 쿼리(Q)와 키(K)의 랜드마크를 계산하여 행렬의 부분집합을 평균화하고, Moore-Penrose 역행렬을 근사하여 어텐션 연산에서의 복잡도를 개선하였다. 또한, Linformer[55]는 셀프 어텐션이 Johnson-Lindenstrauss 보조 정리에 의해 저 랭크라는 것을 증명하고, 시퀀스 길이에 대하여 축소된 k 차원으로 투영하여 시퀀스 정보가 혼합되는 것을 유도하였다. 하지만, 아무리 선형 복잡도를 가진 어텐션 블록을 적용하더라도 입력 데이터의 크기가 커지거나 트랜스포머의 레이어가 깊어질수록 어텐션 블록이 차지하는 연산 및 메모리 비용이 비례하게 커진다는 단점은 여전히 존재한다.

합성곱 신경망의 장점을 통해 비전 트랜스포머의 효율성을 개선하고자 기존 연구에서는 트랜스포머와 합성곱 신경망을 다양한 방식으로 결합하는 하이브리드 네트워크 구조들을 제안하였다. 예를 들어, CVT[50] 모델은 입력 시퀀스에 대하여 단순한 선형 변환을 시키지 않고, 합성곱 연산을 수행하였고, 또 다른 모델인 LeViT[51]는 모델 하위 레이어에 합성곱 연산을 추가하여 지역적 특징 맵을 충분히 추출한 뒤, 해당 특징 맵들에 대하여 트랜스포머 연산을 수행하였다. 또한, MobileViT[52] 모델은 합성곱 연산과 트랜스포머 구조를 반복적으로 번갈아 사용함으로써 지역적이고 전역적인 특징 정보가 혼합될 수 있도록 유도하였다. 이러한 각각의 네트워크 구조들은 동일한 인코더를 여러 레이어로 쌓아 사용하는 기존 비전 트랜스포머의 형태에서 벗어나, 레이어의 깊이에 따라 여러 단계 (Stage)를 구분 짓고, 각 단계를 서로 다른 형태의 구조로 대체한 네트워크를 제안하였다. 합성곱 신경망을 활용하여 지역적인 문맥 정보를 특징 맵에 포함시키고 인코더의 입력 데이터의 크기를 줄임으로써 비전 트랜스포머의 모델 경량화 및 성능 개선도 이루었다. 하지만, 이러한 개선된 하이브리드 구조들도 어텐션 블록 자체의 연산 비용을 최적화하기 위한 기법은 따로 제안하고 있지 않는다.

Ⅲ. 제안하는 방법

본 논문에서는 그림 2와 같이 비전 트랜스포머의 인코더 내부 멀티 헤드 어텐션 (MHA) 블록에 대해 CSP 네트워크 형태를 결합하여, 특징 맵 채널을 k 비율로 분할한 후 해당 부분에 어텐션 연산을 적용하고, 남은 $(1-k)$ 비율의 채널과 병합하는 방식을 제안한다. 제안하는 방법은 기존 연구의 문제점이었던 입력 데이터의 크기에 대한 어텐션 블록에 부담되는 비용 의존도를 낮추기 위해, 어텐션 연산을 수행할 입력 데이터의 채널의 크기를 제한하는 구조를 갖는

다. 하지만, 제안하는 방법을 모델의 모든 어텐션 블록에 적용하였을 때, 레이어의 깊이에 따라 정확도 손실에 대한 민감도가 다를 수 있다. 따라서 레이어 별 고유 민감도를 실험적으로 프로파일링하고, 제안하는 방법을 부분적으로 적용함으로써 상충 관계에 있는 어텐션 블록의 네트워크 복잡도와 모델의 정확도를 최적화하고자 한다. 또한, 제안하는 방법의 범용성과 확장 가능성을 평가 및 검증하기 위해 다양한 선형 어텐션 모델들에 독립적으로 적용하였고, 개선된 하이브리드 구조에 대해서도 동일한 방식으로 제안 방법을 적용하고 평가하였다.

3.1절에서는 제안하는 방법의 구조와 그 효과를 설명하고, 3.2절에서는 정확도 손실을 방어하기 위해 하위 레이어에 대한 부분적 적용 기법에 대해 소개한다. 그리고 3.3절에서는 트랜스포머 부류 모델에 대한 제안하는 방법의 범용성과 확장 가능성에 대해 설명한다.

1. 채널 비율 조정에 따른 어텐션 블록의 네트워크 비용 절감

제안하는 방법은 그림 2와 그림 3과 같이 멀티 헤드 어텐션 연산을 수행할 특징 맵의 채널 비율을 결정하는 하이퍼파라미터 $k(0 < k \leq 1)$에 따라 채널 축을 기준으로 두 개의 부분으로 나눈다. 이후 어텐션 연산 이후에 이들을 다시 합치는 구조를 취함으로써 셀프 어텐션 블록에서 발생하는 연산량을 k^2으로 줄인다. 이때, 제안하는 방법을 적용한 셀프 어텐션 블록의 계산 복잡도는 k^2 배만큼 감소하고, 채널 분할 비율인 k 값에 따라 학습되는 가중치 행렬을 구성하는 파라미터들의 양도 감소한다. 또한, 학습 시 중복된 그래디언트 (Gradient) 정보의 학습을 예방하여 모델의 학습 효과를 높임으로써[59], 결국 모델의 높은 정확도를 유지하면서 학습에 필요한 epoch 수를 줄일 수 있다. 즉, 제안 방법을 적용하면 연산량과 네트워크 파라미터 수를 줄일 수 있을 뿐만 아니라, 학습 속도와 추론 시간을 개선할 수 있다.

그림 2의 제안하는 어텐션 방식은 적용할 레이어의 개수에 따라 연산량과 네트워크 파라미터, 추론 시간이 선형적인 감소 폭을 갖게 된다. 따라서 제안 방식을 적용할 레이어의 개수에 따른 경량화 정도에 대해 이론적인 예측이 가능하다.

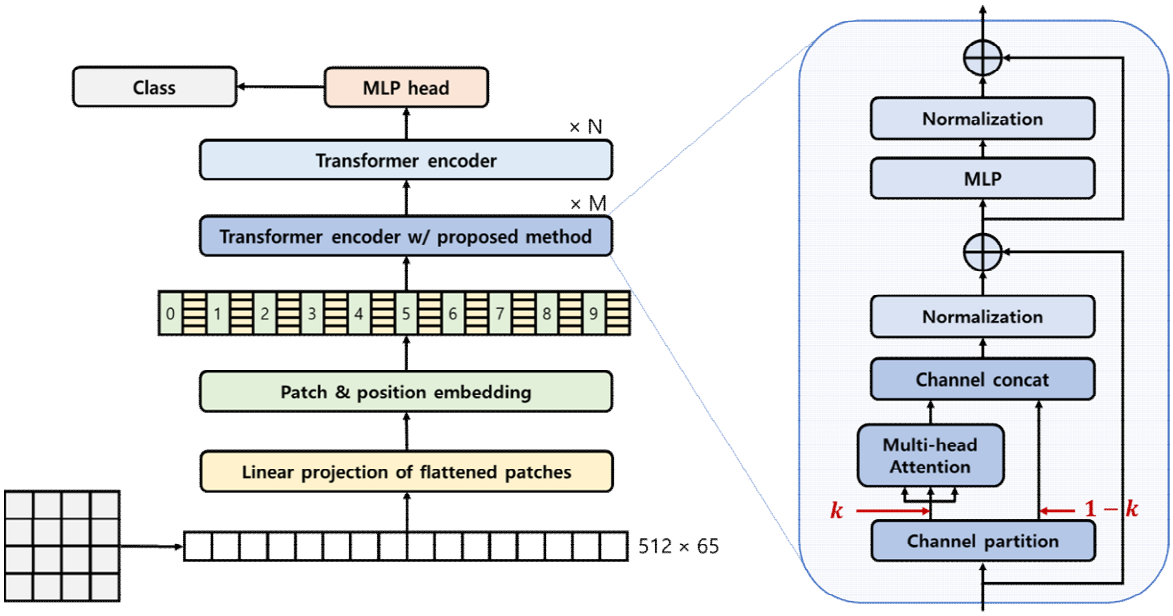


그림 2. 제안하는 적응적 채널 분할 방법을 통한 저복잡도 비전 트랜스포머 (Vision ACPformer)의 블록 다이어그램
Fig. 2. A block diagram representing the proposed low-complexity Vision Transformer with adaptive channel partitioning method (Vision ACPformer)

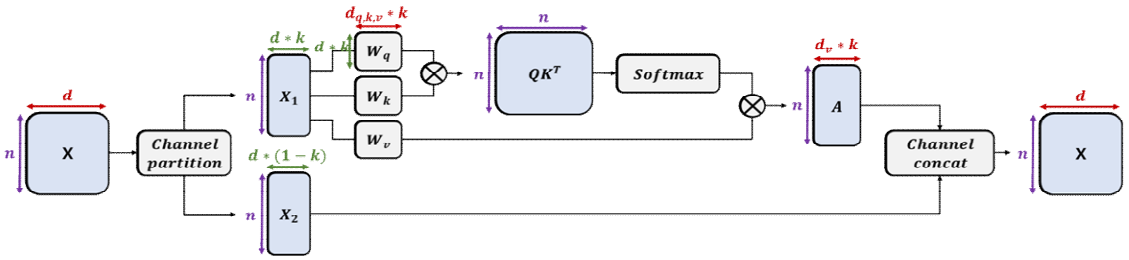


그림 3. 제안하는 적응적 채널 분할 방법의 어텐션 메커니즘
Fig. 3. The proposed attention mechanism using adaptive channel partitioning method

2. 정확도 손실을 최소화하기 위한 하위 레이어 부분적 적용 기법

제안하는 방법을 기존 모델의 특정 레이어에 적용 시 발생할 수 있는 정확도 손실율($Loss_{layer,k}$)은 수식 2와 같이 정의할 수 있다. 해당 수식에서 X_{test}는 평가된 테스트 집합을 의미하며, Acc_{base}는 제안 방법을 적용하지 않았을 때의 분류 정확도이고, $ACC_{layer,k}$는 제안 방법을 하위 첫 번째 레이어에서 하위 $layer$ 번째 레이어까지 k 비율의 채널을 분할했을 때의 분류 정확도를 나타낸다. 이때 민감도($sensitivity_{layer,k}$)는 분류 정확도 손실율에 비례하며, 레이어의 특성이 갖는 고유의 비례 상수 α에 대하여 수식 3과 같이 정의할 수 있다. 수식 3의 정의에 따라, 민감도는 해당 레이어에서 제안 방법을 적용했을 때의 정확도 손실율을 통해 계산되며, 이렇게 정의된 민감도는 어떤 레이어가 제안 방법에 의해 더 큰 영향을 받는지 평가하는 데 사용될 수 있다. 본 논문에서는 실제 실험의 관측 값에 근거하여 레이어 별 민감도를 도출하고자 한다.

$$Loss_{layer,\ k}(X_{test}) = \frac{Acc_{base}(X_{test}) - Acc_{layer,\ k}(X_{test})}{Acc_{base}(X_{test})} \quad (2)$$

$$Sensitivity_{layer,\ k}\ (X_{test}) = \alpha \times Loss_{layer,\ k}(X_{test}) \quad (3)$$

비전 트랜스포머 모델 구조에서 제안하는 적응적 채널 분할 방법을 모든 인코더에 대해 적용할 경우 정확도 손실이 발생할 수 있다. 따라서 잠재적인 정확도 손실을 최소화하기 위해 민감한 레이어에 대해서는 제안 방법을 적용하지 않는 구조를 채택하고자 한다. 비전 트랜스포머의 최종 분류 정확도에 기반하여 레이어 별 민감도를 프로파일링한 결과, 상위 레이어 인코더의 민감도가 하위 레이어 인코더에 비해 더 높은 것을 확인하였다. 따라서 정확도 손실을 최대한 방어하기 위해 상위 레이어 인코더와 하위 레이어 인코더를 구분하고, 하위 레이어 인코더 일부에 대하여 제안하는 어텐션 구조를 부분적으로 적용하였다. 이는 모델의 이미지 분류 작업에 사용되는 특징 맵 결정을 위해 고차원적이고 추상적인 정보를 종합하는 상위 레이어가 에지, 질감 등의 시각 특징들을 추출하는 하위 레이어보다 분류 정확도에 더 직접적인 영향을 미친다고 볼 수 있다. 결과적으로 분류에 중요한 특징 정보를 상위 레이어가 포착하기 위해서는 포괄적인 수용 영역이 필요하며, 이는 토큰들의 관계 정보를 추출하는 어텐션 맵의 해상도가 상대적으로 더 커야 함을 의미한다. 또한, 제안하는 방법을 하위 레이어에만 비대칭적으로 적용함으로써, 정보 집약적인 특징을 추출하는 하위 레이어의 어텐션 블록에서 중복적인 특징 추출이 방지되고, 필수적인 정보만 효율적으로 추출되도록 유도할 수 있다. 이러한 접근 방식을 통해 이미지 분류 응용에서는 정확도 향상 효과를 가져올 수 있다.

3. 제안하는 방법의 범용성 및 확장 적용 가능성

제안하는 어텐션 방식은 독립적인 사용도 가능하지만, 기존의 다양한 어텐션 경량화 방식과 결합하여 동시에 적용하는 것도 가능하다. 이에 더해, 다른 어텐션 블록 경량화 방법들에 비해 직관적이기 때문에, 컴퓨터 비전 응용의 다양한 ViT 기반의 모델들로도 확장하여 적용하기에도 용이하다는 장점을 지닌다. 본 논문에서는 Vanilla Vision Transformer의 어텐션 블록에 대해 적응적 채널 분할 방법

을 적용하였고, 이를 Vision Nystromformer, Vision Performer, Vision Linformer에 대해서도 동일하게 적용하였다. 그리고 개선된 비전 트랜스포머 부류 모델인 CVT, LeViT, MobileViT에 대해서도 제안하는 방법을 채택하였다. 하이브리드 구조의 네트워크에서 입력 데이터의 채널 수와 출력 데이터의 채널 수가 다른 경우에는, 어텐션 연산 이후 병합하는 과정에서 어텐션 연산을 수행하지 않는 채널에 대한 특징 맵 해상도의 조정이 필요하다. 이러한 특징 맵 해상도 불일치 문제를 해결하는 하나의 예시 방법으로, 1×1 합성곱 연산을 포함하는 변형된 적응적 채널 분할 방법을 사용한다.

Ⅳ. 실험 결과

제안하는 방법은 다양한 응용 분야 중 하나의 예시 분야로, 이미지 분류 응용에서 모델의 유효성을 실험적으로 검증하였다. 본 장에서는 비전 트랜스포머 부류의 모델에 제안하는 인코더 구조를 적용할 때, 분류 정확도와 모델 복잡도를 고려하여 최적의 모델 구성을 찾기 위해 수행한 실험 결과를 정리하였다. 표 1과 같이 실험 환경을 구축하고, 비전 트랜스포머에 대한 실험으로 CIFAR10 데이터셋을 사용하고 하이브리드 비전 트랜스포머에 대한 실험으로 Tiny-ImageNet 데이터셋을 사용한다[60]. CIFAR10은 10가지 클래스를 가진 32×32 크기의 해상도의 이미지 데이터셋으로 클래스 당 5000개의 훈련 이미지와 1000개의 테스트 이미지로 구성되어 있다. Tiny-ImageNet 데이터셋은 그보다 다양한 200가지 클래스의 64×64 해상도를 가진 이미지 데이터셋으로 클래스 당 500개의 훈련 이미지와 50개의 검증 이미지, 50개의 테스트 이미지로 구성되어 있다.

표 1. 실험 환경 구성
Table 1. Experimental environment configuration

Components	Products
CPU	Intel® Core™ i9-10940X CPU @ 3.30GHz
Memory	Samsung DDR4 25600 32GB (×8)
GPU	Nvidia Quadro RTX 5000 (×2)

본 실험에서 CIFAR10 이미지 데이터셋 분류를 위한 비전 트랜스포머는 패치 가로 및 세로 크기 4개, 헤드 8개, 임베딩 차원 512개, 멀티 레이어 퍼셉트론 차원 512개, 레이어 6개로 구성된다. 또한, Tiny-ImageNet 데이터셋 분류 실험을 위해 CVT 모델은 세 가지 Stage에 대해 각각 1개의 인코더 블록, 헤드 4개, 임베딩 차원 128개, 퍼셉트론 차원 256개로 구성한다. 또한, LeViT 모델은 Stage 수 1개, 헤드 8개, 임베딩 차원 128개, 퍼셉트론 차원 256개, 레이어 4개로 구성하고, MobileViT는 트랜스포머를 사용하는 세 가지 Stage에 대해 MobileViT 블록 2개, 4개, 3개와 임베딩 차원 96개, 120개, 144개로 구성한다. 배치 크기는 학습 및 추론을 위해 128개로 동일하게 실험하였고, 모든 실험은 5-겹 교차 검증을 통해 진행되었다.

4.1절에서는 Vanilla vision transformer에서 채널 분할 비율을 다르게 조정하여 실험하였다. 또한, Vanilla vision transformer에서 멀티 헤드 어텐션 블록이 제안하는 채널 분할 방법의 타겟 블록으로써 적절한지 검증하기 위해, 멀티 헤드 어텐션 블록뿐만 아니라 인코더 전체와 멀티 레이어 퍼셉트론 블록에도 적용하고 비교하였다. 4.2절에서는 선형 어텐션 모델 중 우수한 성능을 보여준 Performer, Nystromformer, Linformer 모델을 선정하였고, 제안하는 방법을 선정한 모델들에 각각 적용했을 때의 분류 정확도와 모델 복잡도를 비교하고 평가하였다. 4.3절에서는 하이브리드 구조인 CVT, LeViT, MobileViT에 적용하였을 때의 결과를 비교하였다. 각 절에서는 서로 상충 관계(trade-off)에 있는 분류 정확도와 네트워크 복잡도에 대한 실험을 통해 적용할 제안하는 모듈의 최적 개수를 찾기 위한 실험을 수행하였다.

1. 기존 비전 트랜스포머의 분류 정확도 및 네트워크 복잡도 평가

표 2는 제안하는 적응적 채널 분할 방법을 어텐션 블록에 적용하는 것이 다른 블록에 적용하는 것보다 효과적인지 검증하기 위해, 트랜스포머를 구성하는 대표적인 블록 단위인 인코더 (Encoder) 블록, 멀티 헤드 어텐션 (MHA) 블록, 멀티 레이어 퍼셉트론 (MLP) 블록의 세 가지 블록에 대한 실험을 수행하였다. CIFAR10 데이터셋을 사용하여

표 2. 제안하는 방법을 통한 기존 비전 트랜스포머의 분류 정확도 및 네트워크 복잡도 (배치 크기=128, 5-겹 교차 검증)
Table 2. Classification accuracy and network complexity of vanilla Vision Transformer using the proposed method (batch size=128, 5-fold cross validation)

Block Unit using the Proposed Method	Number of Applied Layers	Top 1 Accuracy (%)	FLOPs (K)	Params (K)
Vanilla Vision Transformer	0	80.44	616,856	9,752
Encoder block	1	80.38	540,046	8,571
	2	77.89	463,236	6,994
	3	77.42	386,426	5,418
	4	76.84	309,616	3,841
	5	72.70	232,805	2,265
	6	70.56	155,995	688
MHA block	1	80.84	565,672	8,965
	2	81.19	514,487	7,388
	3	81.62	463,303	5,812
	4	81.24	412,118	4,236
	5	80.01	360,933	2,659
	6	78.83	309,749	1,083
MLP block	1	78.97	591,230	9,358
	2	78.17	565,605	7,781
	3	77.93	539,979	6,205
	4	77.66	514,354	4,628
	5	76.72	488,728	3,052
	6	76.39	463,102	1,476

비전 트랜스포머에 제안하는 방법을 적용하였을 때의 정량적 지표를 정리하였으며, 제안하는 방법이 적용된 블록마다 가장 높은 분류 정확도를 나타내는 지표는 붉은색으로 표기하였다. 그 결과, 인코더 또는 멀티 레이어 퍼셉트론 블록에 제안하는 방법을 적용한 경우, 분류 정확도가 지속적으로 감소한 데 반해 멀티 헤드 어텐션 블록에 적용한 경우에는 기존 비전 트랜스포머와 동등하거나 더 높은 분류 정확도를 나타내기도 하였다. 표 2에 따라 멀티 헤드 어텐션 블록의 하위 3개의 레이어에 채널 분할 방법을 적용하는 것을 제안한다. 이는 기존 비전 트랜스포머에 비해 Top 1 Accuracy 기준 약 1.183%p 향상되었고, FLOPs 약 24.9%, 네트워크 파라미터 수 약 40.4%가 개선되었다. 또한, 배치 크기 128 기준 추론 시간은 16.7% 줄어들었다.

표 3은 제안하는 방법에서 분할할 채널의 비율인 k 값을

표 3. 채널 분할 비율 k에 대한 따른 비전 트랜스포머의 분류 정확도 및 네트워크 복잡도 지표 (배치 크기=128, 5-겹 교차 검증)
Table 3. Classification accuracy and network complexity of vanilla Vision Transformer according to channel partitioning ratio (batch size=128, 5-fold cross validation)

Channel Partitioning Ratio k	Number of Applied Layers	Top 1 Accuracy (%)	FLOPs (K)	Inference time (ms/f)	Params (K)
1	0	80.44	616,856	56.72	9,752
1 / 4	1	80.79	587,004	55.80	9,293
	2	81.59	557,152	54.97	7,716
	3	81.33	527,300	52.48	6,140
	4	79.94	497,448	51.40	4,564
	5	79.91	467,596	50.69	2,987
	6	79.22	437,744	48.27	1,411
2 / 4	1	80.84	565,672	53.48	8,965
	2	81.19	514,487	50.71	7,388
	3	81.62	463,303	48.39	5,812
	4	81.24	412,118	46.92	4,236
	5	80.01	360,933	43.50	2,659
	6	78.83	309,749	40.32	1,083
3 / 4	1	80.74	552,859	52.83	8,768
	2	81.16	488,862	48.89	7,191
	3	81.32	424,864	45.09	5,615
	4	79.57	360,867	41.45	4,039
	5	78.76	296,869	37.08	2,462
	6	75.69	232,872	32.70	886

다르게 조정하여 실험한 결과를 정리한 표이다. 기존 비전 트랜스포머 모델에 대해 k 값이 $\frac{1}{4}$일 때 하위 레이어 3개에 제안하는 방법을 적용한 것이 분류 정확도가 가장 높게 나타났으며, FLOPs 기준 약 31.1%, 추론 시간 약 21.5%, 네트워크 파라미터 수 약 42.4% 감소하였다. 하이퍼파라미터 k 값이 $\frac{1}{2}$일 때는 표 2의 결과와 같으며, $\frac{3}{4}$일 때는 하위 레이어 2개에 제안하는 방법을 적용한 것이 분류 정확도가 가장 높게 나타났다. 이때, 경량화 효과는 FLOPs 기준 약 9.7%, 추론 시간 약 3.1%, 네트워크 파라미터 수 약 21.9% 줄어들었다. 표 3을 통해 제안 방법을 적용하는 채널의 비율과 레이어의 개수에 따라 FLOPs, 추론 시간, 네트워크 파라미터 수가 선형적으로 감소하는 것을 확인할 수 있다. 분류 정확도는 각각의 k 값에 따라서 각각 3개, 3개, 2개의 하위 레이어에 제안하는 방법을 적용했을 때, 기존 비전 트랜스포머에 비해 Top 1 Accuracy 기준 약 0.88%p, 1.18%p, 1.15%p 증가하였다.

2. 선형 어텐션 기반 비전 트랜스포머의 분류 정확도 및 네트워크 복잡도 평가

표 4는 선형적인 비용을 갖는 어텐션 방식인 Performer, Nystromformer, Linformer를 채택한 비전 트랜스포머에 대하여, 제안하는 적응적 채널 분할 방법을 적용했을 때의 객관적 지표들을 나타낸 표이다. 각 모델에 대하여 제안하는 방법을 적용하기 전에는 세 가지 모델 중 Vision Nystromformer가 Top 1 Accuracy 82.96%p로 가장 높은 Top 1 Accuracy를 달성하였으며, 그 다음으로 Vision Linformer가 79.88%p로 높았고, Vision Performer가 78.45%p로 가장 낮았다. 이때 각 모델에 제안하는 방법을 적용하여 실험한 결과 표 4와 같이 Vision Performer, Vision Nystromformer, Vision Linformer에 대하여 각각 레이어 3개, 3개, 4개에 적용했을 때 가장 높은 분류 정확도를 얻었으며, 분류 정확도는 2.33%p, 3.34%p, 1.39%p 향상되었다. 또한, 모델의 경량화 효과로는 각각에 대해서 FLOPs 약 29.8%, 24.9%, 25.6%가 감소하였고, 네트워크 파라미터는 약 41.5%, 41.5%, 56.1% 감소하였다.

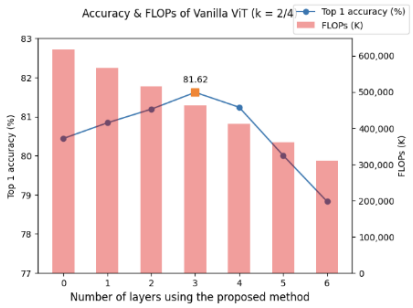
그래프 1은 기존 비전 트랜스포머와 세 가지 선형 어텐션 기반 비전 트랜스포머 모델에 대해 제안하는 방법이 적용된 하위 레이어 개수에 따른 분류 정확도를 그래프로 그린 것이다. 그래프 1과 같이 제안하는 방법을 적용하였을 때 레이어 별로 민감도가 다르기 때문에 정확도 손실율이 다르게 나타날 수 있다. 따라서 그래프 1에 근거하여, 제안하는 방법을 기존 비전 트랜스포머의 하위 레이어 3개

표 4. 비전 나이스트롬포머, 비전 퍼포머, 비전 린포머에 대한 분류 정확도 및 네트워크 복잡도 측정 결과
Table 4. Classification accuracy and network complexity of Vision Performer, Vision Nystromformer, and Vision Linformer

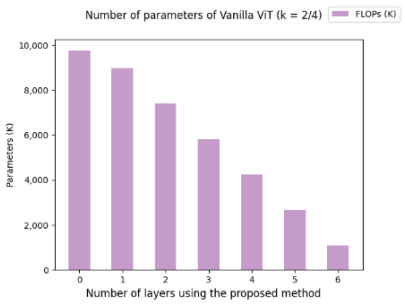
Model	Number of Applied Layers	Top 1 Accuracy (%)	FLOPs (K)	Params (K)
Vision Performer[53]	0	78.45	616,747	9,501
	1	78.63	565,553	8,712
	2	81.09	514,368	7,134
	3	**81.79**	463,183	5,556
	4	80.74	411,999	3,978
	5	78.35	360,814	2,401
	6	76.55	309,629	822
Vision Nystromformer[54]	0	82.96	1,026,075	9,493
	1	83.58	924,264	8,705
	2	84.94	822,453	7,129
	3	**85.29**	720,642	5,552
	4	84.91	618,830	3,975
	5	81.48	517,019	2,399
	6	80.37	415,208	821
Vision Linformer[55]	0	79.88	425,044	6,542
	1	80.63	397,822	6,123
	2	80.86	370,598	5,038
	3	81.04	343,375	3,954
	4	**81.27**	316,152	2,869
	5	78.97	288,929	1,784
	6	78.42	261,706	699

그래프 1. 제안 방법 적용 레이어 별 분류 정확도 및 네트워크 복잡도 (a), (b) 기존 비전 트랜스포머 실험 결과 (c), (d) 비전 퍼포머 실험 결과 (e), (f) 비전 나이스트롬포머 실험 결과 (g), (h) 비전 린포머 실험 결과

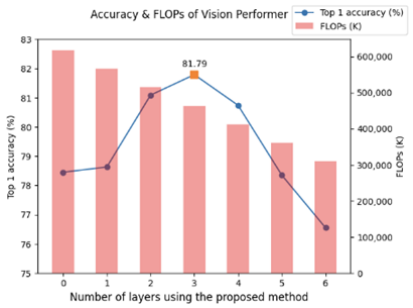
Graph 1. Clasification accuracy and network complexity according to the number of bottom-located layers with the proposed method (a), (b) Results for Vanilla Vision Transformer (c), (d) Results for Vision Performer (e), (f) Results for Vision Nystromformer (g), (h) Results for Vision Linformer



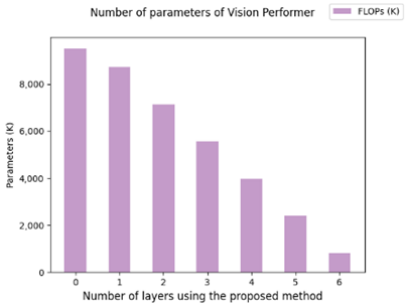
(a)



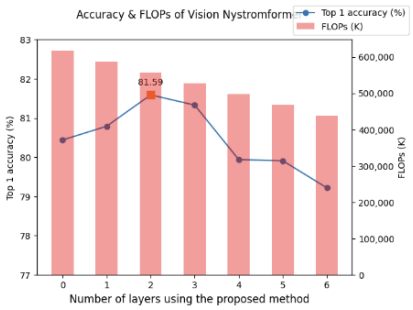
(b)



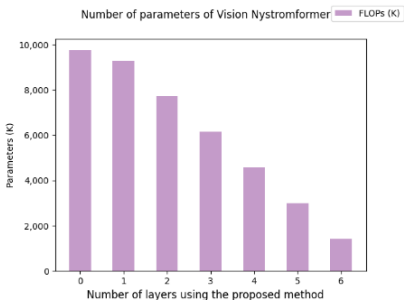
(c)



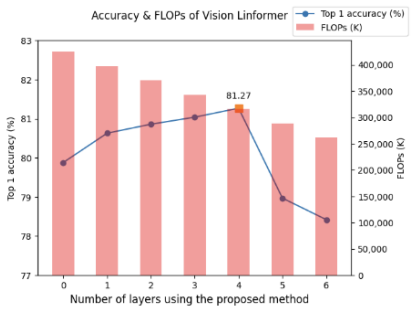
(d)



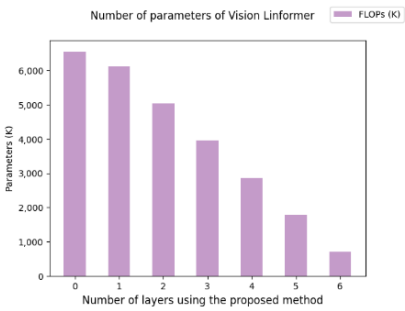
(e)



(f)



(g)



(h)

에 적용하고, Vision Performer, Vision Nystromformer, VisionLinformer 각각에 대해서는 하위 레이어 3개, 3개, 4개에 대해 적용하고자 한다.

3. 하이브리드 구조 기반 비전 트랜스포머의 분류 정확도 및 네트워크 복잡도 평가

표 5는 합성곱 신경망과 트랜스포머 구조를 함께 활용한 하이브리드 구조 기반의 비전 트랜스포머 모델에 대해 객관적 지표를 비교한 표이다. 실험 데이터셋으로 Tiny-ImageNet을 사용했으며, 제안하는 방법을 적용하지 않은 상태에서는 CVT 모델이 Top 1 Accuracy와 Top 5 Accuracy 모두에서 높은 성능을 달성하였다. 각각에 대해 채널 분할 비율을 $\frac{1}{2}$로 설정하고 제안하는 방법을 적용하였을 때, CVT 모델은 하위 2개의 레이어에 적용한 것이 경량화 효과가 높았고, LeViT 모델은 최하위 레이어 1개에 적용하였을 때 가장 높은 분류 정확도를 기록하였다. MobileViT 모델의 경우 제안하는 방법을 4개 또는 5개의 하위 레이어에 적용한 것이 분류 정확도가 높았다.

V. 결 론

본 논문에서는 비전 트랜스포머의 네트워크 복잡도를 효율적으로 개선하는 적응적 채널 분할 방법을 제안하였다. 데이터셋으로는 CIFAR10과 Tiny-ImageNet을 사용하였고, 하나의 응용 예시로써 이미지 분류를 위한 실험을 5-겹 교차 검증 방법으로 수행하였다. 제안하는 모델 구조의 분류 정확도를 판단하는 기준으로 Top 1 및 Top 5 Accuracy 지표를 사용하고, 네트워크의 복잡성을 평가하는 기준으로 부동 소수점 연산 횟수, 추론 시간, 네트워크 파라미터를 사용하여 비교하였다.

실험 결과, 채널 분할 비율을 절반으로 설정하였을 때, 비전 트랜스포머의 앞 3개의 인코더 레이어에 제안하는 방법을 적용하여 FLOPs 약 24.9%, 파라미터 수 약 40.4%, 추론 속도 약 16.7%를 개선하였으며, 평균 정밀도(Average Precision, AP) 기준 1.183%p 분류 정확도를 향상시켰다. 또한, Vision Performer와 Vision Nystromformer에 대해서는 3개의 레이어에 제안하는 방법을 적용한 모델을, Vision Linformer에 대해서는 4개의 레이어에 적용한 모델을 최적의 모델로 선정하였다. 그리고 개선된 하이브리드

표 5. 하이브리드 비전 트랜스포머 모델 별 제안하는 방법 적용 시 분류 정확도 및 네트워크 복잡도
Table 5. Classification accuracy and network complexity of hybrid Vision Transformer using the proposed method

Model	Number of Applied Layers	Top 1 Accuracy (%)	Top 5 Accuracy (%)	FLOPs (K)	Params
CVT[50]	0	41.98	69.03	842,556	850,888
	1	41.61	68.54	792,249	801,672
	2	**41.81**	**68.73**	741,893	752,456
	3	40.58	67.84	691,596	703,240
	4	40.05	67.35	641,299	654,024
LeViT[51]	0	32.96	60.32	632,284	645,384
	1	**34.73**	**62.95**	581,953	596,168
	2	33.84	60.97	531,621	546,952
	3	29.01	55.88	481,289	497,736
	4	28.49	53.95	430,958	448,520
MobileViT[52]	0	31.55	58.89	115,796	2,075,744
	1	32.04	59.32	114,592	2,066,384
	2	32.33	58.80	113,388	2,016,464
	3	32.53	59.53	57,108	2,054,684
	4	**33.92**	60.39	56,318	1,873,004
	5	33.62	**61.21**	56,130	1,741,244
	6	32.50	60.21	55,941	1,609,484
	7	31.93	59.30	55,885	1,595,444
	8	31.79	58.83	55,828	1,409,684
	9	31.22	57.33	55,772	1,223,924

구조인 CVT, LeViT, MobileViT에 대해서는 제안하는 방법을 각각 2개, 1개, 4개의 하위 레이어에 적용하였을 때, 본래의 높은 분류 정확도를 유지하는 모습을 보여주었다. 본 논문에서 제안하는 방법을 통해 향후 트랜스포머 기반 모델의 복잡도를 줄이는데 도움을 줄 것으로 기대된다. 또한, 이미지 분류 분야 뿐만 아니라, 이미지 세그먼테이션과 이미지 복원과 같은 다른 응용 분야에서도 객체의 세부적인 텍스처와 경계선의 특성을 포착하고 세밀한 픽셀 단위의 정보를 복원하는 데 적용이 용이할 것으로 보인다.

참 고 문 헌 (References)

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
doi: https://doi.org/10.48550/arXiv.1706.03762

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.
doi: https://doi.org/10.48550/arXiv.2010.11929

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” Association for Computational Linguistics, vol. 1, June 2019.
doi: https://doi.org/10.18653/v1/N19-1423

[4] B. L. Edelman, S. Goel, S. Kakade, and C. Zhang, “Inductive Biases and Variable Creation in Self-Attention Mechanisms,” in Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, pp. 5793－5831, 2022.

[5] S. Garg, D. Tsipras, P. S. Liang, and G. Valiant, “What can transformers learn in-context? a case study of simple function classes,” Advances in Neural Information Processing Systems, vol. 35, pp. 30583-30598, 2022.

[6] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in International conference on machine learning, PMLR, pp. 4055-4064, 2018.

[7] K. Islam, “Recent advances in vision transformer: A survey and outlook of recent work,” arXiv preprint arXiv:2203.01536, 2022.
doi: https://doi.org/10.48550/arXiv.2203.01536

[8] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” ACM computing surveys (CSUR), vol. 54, no. 10s, pp. 1-41, 2022.
doi: https://doi.org/10.1145/3505244

[9] S. Jamil, M. Jalil Piran, and O.-J. Kwon, “A comprehensive survey of transformers for computer vision,” Drones, vol. 7, no. 5, pp. 287, 2023.
doi: https://doi.org/10.3390/drones7050287

[10] J. Kugarajeevan, T. Kokul, A. Ramanan, and S. Fernando, “Transformers in single object tracking: An experimental survey,” IEEE Access, 2023.
doi: https://doi.org/10.1109/ACCESS.2023.3298440

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European conference on computer vision, Springer, pp. 213-229, 2020.
doi: https://doi.org/10.1007/978-3-030-58452-8_13

[12] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 357-366, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00041

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” arXiv preprint arXiv:2102.04306, 2021.
doi: https://doi.org/10.48550/arXiv.2102.04306

[14] Y. Lee, E Lee, M Lee, J Byeon, H Ahn, and D Sim, "Deep Learning-Based Image Enhancement Techniques for Maritime Video in Storage and Transmission Systems: A Research Study," Journal of Broadcast Engineering, Vol.28, No.4, pp.410-428, July 2023.
doi: https://doi.org/10.5909/JBE.2023.28.4.410

[15] A. M. Ali, B. Benjdira, A. Koubaa, W. El-Shafai, Z. Khan, and W. Boulila, "Vision transformers in image restoration: A survey," Sensors, vol. 23, no. 5, pp. 2385, 2023.
doi: https://doi.org/10.3390/s23052385

[16] A. B. Koyuncu, H. Gao, A. Boev, G. Gaikov, E. Alshina, and E. Steinbach, "Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression," in European Conference on Computer Vision, Springer, pp. 447-463, 2022.
doi: https://doi.org/10.1007/978-3-031-19800-7_26

[17] J. Lee, J. Park, H. Choi, J. Byeon, D. Sim, “Overview of VVC,” Broadcasting and Media Magazine, Vol.24, No.4, pp.10-25, Apr 2019.

[18] M. Lee, H. Song, J. Park, B. Jeon, J. Kang, J.-G. Kim, Y. Lee, J.-W. Kang, and D. Sim, “Overview of Versatile Video Coding (H. 266/VVC) and Its Coding Performance Analysis,” IEIE Transactions on Smart Processing & Computing, Vol.12, No.2, pp.122-154, Apr 2023.
doi: https://doi.org/10.5573/IEIESPC.2023.12.2.122

[19] X. Wang, L. L. Zhang, Y. Wang, and M. Yang, "Towards efficient vision transformer inference: A first study of transformers on mobile devices," in Proceedings of the 23rd annual international workshop on mobile computing systems and applications, pp. 1-7, 2022.
doi: https://doi.org/10.1145/3508396.3512869

[20] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, “Transformers in time series: A survey,” arXiv preprint arXiv:2202.07125, 2022.
doi: https://doi.org/10.48550/arXiv.2202.07125

[21] F. D. Keles, P. M. Wijewardena, and C. Hegde, "On the computational complexity of self-attention," in International Conference on Algorithmic Learning Theory, PMLR, pp. 597-619, 2023.

[22] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," ACM Computing Surveys, vol. 55, no. 6, pp.

1-28, 2022.
doi: https://doi.org/10.1145/3530811

[23] J. Pan, A. Bulat, F. Tan, X. Zhu, L. Dudziak, H. Li, G. Tzimiropoulos, and B. Martinez, "Edgevits: Competing light-weight cnns on mobile devices with vision transformers," in European Conference on Computer Vision, Springer, pp. 294-311, 2022.
doi: https://doi.org/10.1007/978-3-031-20083-0_18

[24] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, R. M. Anwer, and F. Shahbaz Khan, "Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications," in European Conference on Computer Vision, Springer, pp. 3-20, 2022.
doi: https://doi.org/10.1007/978-3-031-25082-8_1

[25] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "SwiftFormer: Efficient additive attention for transformer-based real-time mobile vision applications," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17425-17436, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.01598

[26] Y. Meng, P. Wu, J. Feng, and X. Zhang, "MixMobileNet: A Mixed Mobile Network for Edge Vision Applications," Electronics, vol. 13, no. 3, p. 519, 2024.
doi: https://doi.org/10.3390/electronics13030519

[27] J. Yang, J. Liao, F. Lei, M. Liu, J. Chen, L. Long, H. Wan, B. Yu, and W. Zhao, "TinyFormer: Efficient Transformer Design and Deployment on Tiny Devices," arXiv preprint arXiv:2311.01759, 2023.
doi: https://doi.org/10.48550/arXiv.2311.01759

[28] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient processing of deep neural networks: A tutorial and survey,” Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, 2017.
doi: https://doi.org/10.1109/JPROC.2017.2761740

[29] J. Park, J. Lee, and D. Sim, “Low-complexity 1D-convolutional Neural Network for Super Resolution,” IEIE Transactions on Smart Processing & Computing, vol. 8, no. 6, pp. 423-430, 2019.
doi: https://doi.org/10.5573/IEIESPC.2019.8.6.423

[30] Y. Wu, D. Wang, X. Lu, F. Yang, G. Li, W. Dong, and J. Shi, “Efficient visual recognition with deep neural networks: A survey on recent advances and new directions,” arXiv preprint arXiv:2108.13055, 2021.
doi: https://doi.org/10.1007/s11633-022-1340-5

[31] J. Lee, L. Mukhanov, A. S. Molahosseini, U. Minhas, Y. Hua, J. M. del Rincon, K. Dichev, C.-H. Hong, and H. Vandierendonck, “Resource-efficient deep learning: A survey on model-, arithmetic-, and implementation-level techniques,” arXiv preprint arXiv:2112. 15131, 2021.
doi: https://doi.org/10.48550/arXiv.2112.15131

[32] M. Capra, B. Bussolino, A. Marchisio, M. Shafique, G. Masera, and M. Martina, “An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks,” Future Internet, vol. 12, no. 7, pp. 113, 2020.
doi: https://doi.org/10.3390/fi12070113

[33] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, “Rethinking the value of network pruning,” arXiv preprint arXiv:1810.05270, 2018.
doi: https://doi.org/10.48550/arXiv.1810.05270

[34] M. Zhu, Y. Tang, and K. Han, “Vision transformer pruning,” arXiv preprint arXiv:2104.08500, 2021.
doi: https://doi.org/10.48550/arXiv.2104.08500

[35] B. Xu, T. Zhang, Y. Wang, and Z. Chen, "A Knowledge-Distillation-Integrated Pruning Method for Vision Transformer," 2022 21st International Symposium on Communications and Information Technologies (ISCIT), September 2022.
doi: https://doi.org/10.1109/iscit55906.2022.9931309

[36] Y. Cao, “FCP_DIS_ViT: Efficient Vision Transformer with Neural Network Pruning,” 2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA), January 2024.
doi: https://doi.org/10.1109/icpeca60615.2024.10470980

[37] W. Hao, P. Judd, X. Zhang, M. Isaev and P. Michikevicius, “Integer Quantization For Deep Leaning Inference: Principles and Empirical Evaluation,” arXiv preprint, April 2020.
doi: https://doi.org/10.48550/arXiv.2004.09602

[38] I. Chung, B. Kim, Y. Choi, S. Kwon, Y. Jeon, B. Park, S. Kim and D. Lee, “Extremely Low Bit Transformer Quantization for On-Device Neural Machine Translation,” Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.
doi: https://doi.org/10.18653/v1/2020.findings-emnlp.433

[39] P. Nayak, D. Zhang, and S. Chai, “Bit Efficient Quantization for Deep Neural Networks,” 2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing – NeurIPS Eddition (EMC2-NIPS), December 2019.
doi: https://doi.org/10.1109/emc2-nips53020.2019.00020

[40] Z. Li and Q. Gu, "I-ViT: integer-only quantization for efficient vision transformer inference," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17065-17075, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.01565

[41] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
doi: https://doi.org/10.48550/arXiv.1503.02531

[42] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
doi: https://doi.org/10.48550/arXiv.1910.01108

[43] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International conference on machine learning, 2021: PMLR, pp. 10347-10357, 2021.

[44] P. Deng, K. Xu, and H. Huang, "When CNNs meet vision transformer: A joint framework for remote sensing scene classification," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2021.
doi: https://doi.org/10.1109/LGRS.2021.3109061

[45] B. K. Iwana and A. Kusuda, "Vision Conformer: Incorporating Convolutions into Vision Transformer Layers," in International Conference on Document Analysis and Recognition, Springer, pp. 54-69, 2023.
doi: https://doi.org/10.1007/978-3-031-41685-9_4

[46] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “Coatnet: Marrying convolution and attention for all data sizes,” Advances in neural information processing systems, vol. 34, pp. 3965-3977, 2021.

[47] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition, pp. 815-825, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00089

[48] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 3286-3295, 2019.
doi: https://doi.org/10.1109/ICCV.2019.00338

[49] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging mobilenet and transformer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5270-5279, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00520

[50] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 22-31, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00009

[51] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 12259-12269, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.01204

[52] S. Mehta, and M. Rastegari, “Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer,” arXiv preprint arXiv:2110.02178, 2021.
doi: https://doi.org/10.48550/arXiv.2110.02178

[53] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, and L. Kaiser, “Rethinking attention with performers,” arXiv preprint arXiv: 2009.14794, 2020.
doi: https://doi.org/10.48550/arXiv.2009.14794

[54] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh, "Nyströmformer: A nyström-based algorithm for approximating self-attention," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp. 14138-14148, 2021.
doi: https://doi.org/10.1609/aaai.v35i16.17664

[55] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: Self-attention with linear complexity,” arXiv preprint arXiv:2006. 04768, 2020.
doi: https://doi.org/10.48550/arXiv.2006.04768

[56] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150, 2020.
doi: https://doi.org/10.48550/arXiv.2004.05150

[57] N. Kitaev, Ł. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” arXiv preprint arXiv:2001.04451, 2020.
doi: https://doi.org/10.48550/arXiv.2001.04451

[58] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, “Efficient content-based sparse attention with routing transformers,” Transactions of the Association for Computational Linguistics, vol. 9, pp. 53-68, 2021.
doi: https://doi.org/10.1162/tacl_a_00353

[59] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 390-391, 2020.
doi: https://doi.org/10.1109/CVPRW50498.2020.00203

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255, 2009.
doi: https://https://doi.org/10.1109/CVPR.2009.5206848

저 자 소 개

이 세 영

- 2022년 8월 : 광운대학교 컴퓨터공학과 학사
- 2022년 9월 ~ 현재 : 광운대학교 컴퓨터공학과 석사
- ORCID : https://orcid.org/0009-0002-2500-7228
- 주관심분야 : 영상신호처리, 영상압축, 컴퓨터비전

오 상 수

- 2024년 2월 : 광운대학교 컴퓨터정보공학부 학사
- ORCID : https://orcid.org/0009-0001-0211-8614
- 주관심분야 : 신호처리, 컴퓨터비전

저 자 소 개

한 상 현

- 2024년 2월 : 광운대학교 컴퓨터정보공학부 학사
- ORCID : https://orcid.org/0009-0007-8654-4533
- 주관심분야 : 컴퓨터비전

임 승 환

- 2024년 2월 : 광운대학교 컴퓨터정보공학부 학사
- ORCID : https://orcid.org/0009-0001-5824-2470
- 주관심분야 : 컴퓨터비전, 딥러닝

이 종 석

- 2016년 2월 : 광운대학교 전자공학과 학사
- 2018년 2월 : 광운대학교 전자공학과 석사
- 2024년 2월 : 광운대학교 컴퓨터공학과 박사
- 2020년 ~ 2024년 : 디지털인사이트 선임연구원
- 2024년 2월 ~ 현재 : LG전자 선임연구원
- ORCID : https://orcid.org/0000-0001-8045-0244
- 주관심분야 : 영상압축, 컴퓨터비전, 고해상도 위성영상처리, 포인트 클라우드 압축

심 동 규

- 1993년 2월 : 서강대학교 전자공학과 공학사
- 1995년 2월 : 서강대학교 전자공학과 공학석사
- 1999년 2월 : 서강대학교 전자공학과 공학박사
- 1999년 3월 ~ 2000년 8월 : 현대전자 선임연구원
- 2000년 9월 ~ 2002년 3월 : 바로비젼 선임연구원
- 2002년 4월 ~ 2005년 2월 : University of Washington Senior research engineer
- 2005년 3월 ~ 현재 : 광운대학교 컴퓨터공학과 교수
- ORCID : https://orcid.org/0000-0002-2794-9932
- 주관심분야 : 영상신호처리, 영상압축, 컴퓨터비전