Efficient Vision Transformer 모델 개발

팀 명 HoViT

팀 원 정용훈, 남현원, 한동엽, 강전찬

지도교수 이상훈

2025-1 AJOU SOFTCON

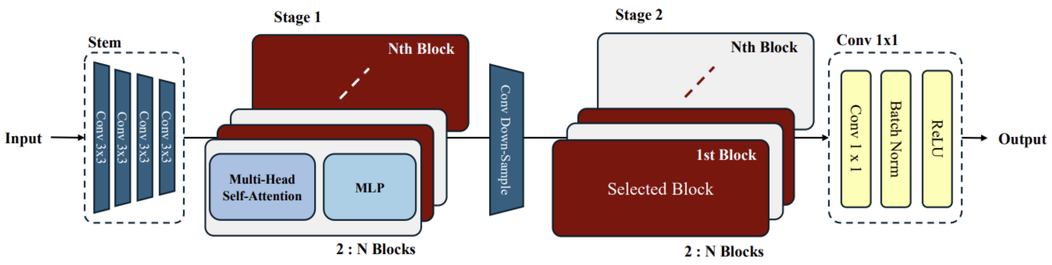
연구배경 및 관련연구

연구 배경 : ViT와 같은 트랜스포머 기반 모델은 우수한 성능을 보이지만 높은 연산 복잡도와 많은 파라미터로 인해 경량화가 필요한 환경에서는 실용성이 떨어진다. 이를 보완하기 위해 ResNet과 같은 CNN이나 LeViT와 같은 경량 트랜스포머 구조가 제안되었으나 여전히 추론 속도 개선과 성능 유지 사이의 균형에는 한계가 있다. 특히 의료 영상 분야에서는 정확도뿐만 아니라 실시간 추론 속도도 중요하기 때문에 이를 해결하기 위한 경량화 및 속도 개선 연구가 활발히 진행되고 있다. 따라서 본 연구에서는 추론 속도를 높이면서도 정확도를 유지할 수 있는 HoViT(Hybrid and Optimized Vision Transformer) 구조를 제안한다.

CNN vs Transformer : CNN은 입력 이미지의 국소 정보를 계층적으로 추출하며 연산 효율이 높아 실시간 처리에 적합하다. 반면, ViT는 Self-Attention을 통해 전역적 관계를 학습할 수 있어 표현력이 뛰어나지만,입력 크기에 따라 연산량이 급증하여 고해상도 이미지 처리에는 비효율적일 수 있다. 이로 인해 두 구조는 각각의 장단점에 따라 활용 분야가 다르게 나타난다.

Efficient Architecture : EfficientNet은 너비, 깊이, 해상도를 균형 있게 조절하는 방식으로 높은 성능과 낮은 연산량을 동시에 달성한 CNN 기반 경량화 모델이다. LeViT는 CNN과 트랜스포머의 장점을 결합한 하이브리드 구조로 CNN으로 특징을 추출한 뒤 효율적인 Attention Block으로 처리함으로써 표현력은 유지하면서 추론 속도와 연산 효율을 높였다.

개발내용



[그림1] Overall architecture of HoViT

Design principle : HoViT는 LeViT에서 불필요한 구성 요소를 제거해 추론 속도를 개선한 최적화 아키텍처로, CNN과 트랜스포머 구조를 결합한다. 초기에는 ResNet 계층을 유지하면서, 각 Attention 블록의 병목을 분석해 효율적으로 축소하였고, 중요한 블록만을 선택적으로 남기는 학습 기반 프루닝 기법을 적용해 계산 효율을 더욱 높였다.

stage block 수 : LeViT 모델에서는 각 스테이지에 3개의 Attention 블록이 존재하여 계산량이 많아 추론 속도에 부담이 되는 문제가 있었다. 이러한 문제를 해결하기 위해 각 스테이지의 Attention 블록 수를 기존 3개에서 2개로 줄이는 방식을 도입하여, Attention 연산의 반복 횟수를 줄여 계산 복잡성을 낮추고 효율성을 개선하였다.

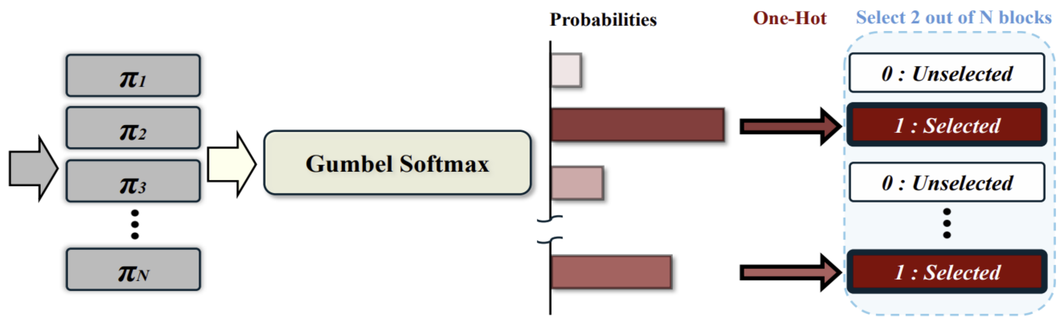
Downsampling 방식 변경 : 기존 LeViT 모델은 Attention 기반의 downsampling을 진행한다. 이 과정은 feature map을 압축하는 역할을 하는데 attention 연산을 사용한다.

$$Attention_{downsample}(Q'_{S=2}, K, V) = Softmax(\frac{Q'_{S=2}K^T}{\sqrt{d_k}})V$$

이는 downsampling 과정에서의 복잡한 관계 학습이 오히려 계산량의 증가로 비효율적일 것이라고 판단했고, 이를 컨볼루션 downsampling 방법으로 변경했다.

$$Conv_{downsample} = Conv(x, Kernel = 3, Stride = 2, Padding = 1)$$

해당 변경으로 Attention 기반 downsampling에 비해 연산량을 줄일 수 있고, 그 결과 성능은 유지한 채 추론 시간은 개선되는 결과를 얻었다.



[그림2] Process of Learnable Pruning

1x1 Conv layer로 대체 : LeViT에서는 총 3개의 stage가 존재하는데, 이 중 마지막 stage를 1x1 컨볼루션을 활용하여 채널 수를 384에서 512로 증가 시켰다. 이를 통해 파라미터 수의 증가폭을 줄인다. 그와 동시에 채널 간의 비선형적 조합을 통해 고차원 특징을 정교하게 표현하며 추론 시간 개선하였다.

Learnable Pruning : TinyFusion에서 영감을 받아 Gumbel-Softmax 기반의 학습 가능한 블록 선택 기법을 도입했다. 입력 x에 대해 블록 집합 $\{B_1, B_2, B_N\}$ 에 대한 확률 분포 π를 학습한다.

$$p_i = \frac{\exp((\log(\pi_i) + G_i)/\tau)}{\sum_{j=1}^{N} \exp((\log(\pi_j) + G_j)/\tau)} \quad (1 \le i \le N)$$

여기서 $G_i \sim Gumble(0,1)$ 은 확률적인 샘플링을 가능하게 하는 Gumble 분포를 따르는 노이즈 항이며 τ 는 선택 과정에서의 무작위성 정도를 조절하는 파라미터이다.

$$x_{t+1} = x_t + \sum_{i=1}^{N} p_i B_i(x_t), \quad \forall i \text{ such that } p_i > 0.$$

stage t에서는 각 블록이 학습된 확률에 따라 가중하여 학습 시에는 여러 블록을 혼합해 사용하고 추론 시에는 하나의 블록만 선택해 실행함으로써 연선 효율을 높인다.

Low-Rank Adaptation (LoRA) : 학습 가능한 프루닝 구조에서 전체 파라미터를 미세 조장하는 것은 계산 비용이 크기 때문에 본 연구에서는 LoRA를 도입해 효율적인 최적화를 수행하였다. LoRA는 기존 가중치를 고정한 채 학습 가능한 저랭크 행렬 A, B를 추가하고 이를 통해 $W_{fine-tuned} = W + aAB$ 형태로 파라미터 업데이트를 수행한다. 마지막 1x1 Conv layer는 분류 전 잠재표현을 향상시키기 위해 학습 가능하게 유지된다.

실험 결과 및 분석

표-1은 Vanilla ViT, ViT-Tiny, LeViT과 같은 기존 Transformer 계열 모델뿐만 아니라 ResNet, EfficientNet과도 성능을 비교한다. HoViT은 모든 평가 지표에서 경쟁력을 유지하면서도 기존 SOTA 모델인 EfficientNet-B0보다 41.4% 더 빠른 추론 속도를 보여준다. 이는 convolution과 attention 을 결합한 하이브리드 구조, 효율적인 block 선택 그리고 학습 가능한 가지치기(pruning)을 통해 이루어진 결과이다.

표에서 제시된 HoViT-B은 데이터 증강(augmentation) 없이, 학습 가능한 가지치기만 적용된 구성으로, 속도와 성능 간의 균형을 가장 효과적으로 달성하는 모델이다. 특히 HoViT은 더 많은 파라미터 수를 가진 LeViT보다 더 높은 정확도와 빠른 추론 속도를 달성하며, 효율적으로 경량화된 구조와 학습 가능한 가지치기를 통해 자원 제한 환경에서도 높은 효율성을 보인다.

표-2는 HoViT의 핵심 설계 요소에 대해 ablation 실험을 수행한다. 이를 통해 성능 및 추론 속도에 미치는 영향을 평가하였다. 실험은 convolution-based downsampling과 학습 가능한 가지치기의 효과를 중심으로 진행되었다.

먼저, attention-based downsampling과 convolution-based downsampling을 비교한 결과, 후자를 적용한 HoViT이 약 3% 더 높은 성능을 보였으며, 추론 속도 또한 소폭 감소하였다. 이는 convolution이 지역적 특징 학습에 효과적이면서도 효율성을 유지함을 보여준다.

또한, 학습 가능한 가지치기의 효과를 분석하기 위해 pruning을 적용하지 않은 HoViT-S와 pruning을 적용한 HoViT-B를 비교하였다. 그 결과, HoViT-B는 더 높은 성능을 유지하면서도 추론 속도 저하 없이 작동하였으며, 이는 특징 표현의 최적화를 효과적으로 달성했음을 시사한다.

표-1. 기존 모델과의 성능 및 추론 시간 비교

Architecture	# params (M)	Performance Top-1	F1	Recall	Precision	Inference time Average (ms)
HoViT-B (ours)	5.0	98.90	98.88	98.92	98.85	6.15
HoViT-B w/o P, w/ Aug	8.3	98.92	98.93	98.94	98.91	10.10
HoViT-B w/o P	8.3	98.68	98.63	98.65	98.61	10.10
HoViT-S w/o P, w/ Aug	4.9	98.65	98.64	98.70	98.59	6.13
HoViT-S w/o P	4.9	98.16	98.15	98.13	98.18	6.13
ViT-Tiny	5.5	97.40	97.32	97.37	97.29	10.01
LeViT-128S	6.6	95.92	95.82	95.87	95.93	12.06
ViT	85.8	93.36	93.27	93.36	93.33	282.60
EfficientNet-B0	4.0	98.97	98.95	98.93	98.98	10.49
EfficientNet-B1	6.5	98.54	98.53	98.53	98.53	15.54
ResNet-50	23.5	98.73	98.75	98.73	98.78	13.29

표-2. HoViT 설계에 대한 Ablation 연구

	# params (M)	Performance Top-1	F1	Recall	Precision	Inference time Average (ms)
HoViT-B (Ours)	5.0	98.90	98.88	98.92	98.85	6.15
w/o P	4.9	98.63	98.59	98.98	99.03	5.74
w/ attnDown	5.79	95.67	95.57	95.67	95.70	6.10

오픈소스 URL

https://github.com/hun9008/HoViT

Paper QR Link