비전 맘바 가속을 위한 스코어 스무딩 기반의 패치 프루닝 기법

이민영, 정기석*
한양대학교

lmo970901@hanyang.ac.kr, *kchung@hanyang.ac.kr

Score-Smoothing-based Patch Pruning for Vision Mamba Acceleration

Min Young Lee, Ki-Seok Chung*
Hanyang University, Seoul, Korea

요 약

본 논문은 비전 맘바(Vision Mamba)의 추론 효율성을 높이기 위한 스코어 스무딩 기반 패치 프루닝 기법을 제안한다. 비전 맘바는 선형 복잡도를 갖지만, 순차적 데이터 처리 특성상 기존의 무작위 토큰 제거 시 심각한 문맥 손실이 발생한다. 이를 해결하기 위해 경량 컨볼루션 신경망(Lightweight ConvNets)으로 패치 중요도를 예측하고, 스코어 스무딩 필터를 적용하여 인접 패치 간의 일관성을 확보함으로써 클러스터 형태의 토큰 보존을 유도한다. 또한, 인코더 진입 전 단 한 번의 프루닝을 수행하는 원스텝 구조를 통해 추론 지연 시간을 단축한다. 실험 결과, ImageNet-1K 데이터셋에서 토큰 40% 제거 시 별도의 재학습 없이도 73.83%의 정확도를 기록하여 기존 맘바 전용 기법 대비 우수한 성능을 입증했다.

Ⅰ. 서 론

최근 비전 트랜스포머(Vision Transformer)는 우수한 성능을 입증했으나, 해상도 증가에 따른 이차적($O(N^2)$) 복잡도로 인해 고해상도 데이터 처리 시 하드웨어 자원에 큰 부담을 준다. 이를 극복하기 위해 선형 복잡도($O(N)$)를 갖는 상태 공간 모델(State Space Model) 기반의 비전 맘바(Vision Mamba)[1]가 주목받고 있다. 그러나 자율주행이나 의료 영상과 같은 고해상도 환경에서는 비전 맘바 역시 절대적인 연산량과 메모리 요구량이 증가하여 실시간 처리에 병목을 초래한다. 따라서 효율적인 추론을 위한 토큰 프루닝 등 경량화 기법의 도입이 필수적이다.

하지만 히든 스테이트(Hidden State)를 순차적으로 갱신하는 비전 맘바의 구조적 특성상, 기존 비전 트랜스포머의 프루닝 기법을 그대로 적용하면 정보 흐름의 단절과 문맥 손실이 발생한다. 이를 해결하기 위해 본 논문에서는 패치 프루닝(Patch Pruning, PaPr)[2]을 베이스로 하되, 인접 패치 간 중요도 일관성을 보강하는 스코어 스무딩(Score Smoothing, SS) 기법을 추가한 구조를 제안한다. 본 논문은 패치 프루닝 및 스코어 스무딩 기법이 비전 맘바 구조에서 유효한 경량화 전략임을 입증하며, 특히 별도의 재학습 없이도 기존 기법 대비 우수한 성능 유지 능력을 보여준다.

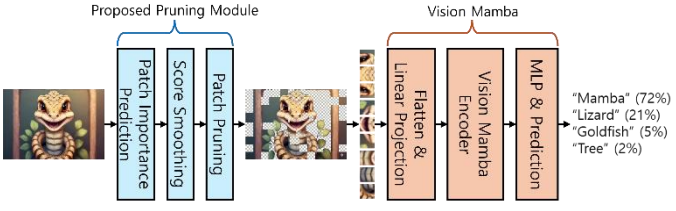

그림 1. 제안하는 전체 시스템의 아키텍처

Ⅱ. 본 론

Ⅱ-1. 이미지 패치 중요도 예측

기존의 프루닝은 모델 중간 레이어의 특징 맵 통계치에 의존하는 것과 달리, 본 연구에서 활용하는 패치 프루닝 기법은 입력 초기 단계에서 패치의 중요도를 결정하는 경량 컨볼루션 신경망(Lightweight ConvNets)[2]을 통해 중요도를 결정한다. 경량 컨볼루션 신경망은 MobileNet-v2 기반의 초경량 구조를 활용하여 입력 이미지로부터 각 패치의 중요도 스코어 맵을 생성한다. 이 과정에서 CNN 기반 추출기는 단순 픽셀 강도를 넘어 주변 컨텍스트를 포함한 의미론적 중요도를 확보하게 된다. 산출된 스코어 맵에 스코어 스무딩을 적용한 후, 상위 K 개의 중요 패치 인덱스를 선택하여 프루닝을 수행한다.

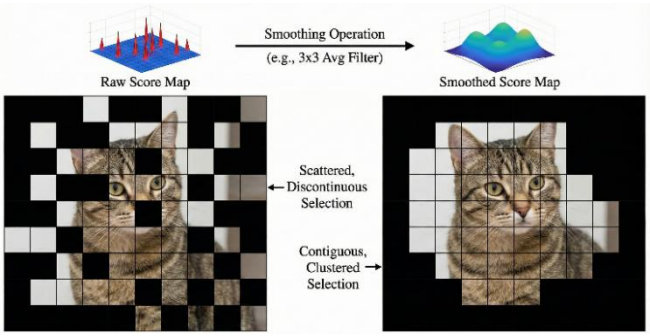

그림 2. 스코어 스무딩 적용 후 패치 프루닝

Ⅱ-2. 스코어 스무딩 (Score Smoothing)

본 논문의 핵심 기여인 스코어 스무딩은 앞선 패치 프루닝 예측 단계에서 생성된 개별 스코어 맵에 공간적 상관관계를 부여하는 필터링 기법이다. 본 기법은 영상 처리의 표준적인 스무딩 원리[3]를 비전 맘바의 스캔 메커니즘이 요구하는 문맥 일관성 보존 목적에 맞게 재해석하여 통합한 것이다.

비전 맘바는 토큰을 순차적으로 스캔하며 히든 스테이트를 갱신하므로, 프루닝 과정에서 토큰들이 무작위로 제거될 경우 상태 전달의 불연속성이 발생하여 정보 손실이 극대화된다. 이를 해결하기 위해, 본 연구에서는 3x3 또는

5x5 크기의 평균 필터를 스코어 맵(S)에 적용하여 보정된 스코어 맵(S_{smooth})을 생성한다. 구체적으로, 위치 i, j 에서의 원래 스코어 $S_{i,j}$ 에 대한 스무딩 연산은 식 (1)과 같이 정의된다.

$$S_{smooth\ i,j} = \frac{1}{k^2} \sum_{(p,q)\in \mathcal{N}_{i,j}} S_{p,q} \quad (1)$$

여기서 k 는 필터의 크기를 나타내며 $N_{i,j}$ 는 i, j 를 중심으로 하는 k x k 영역 내의 패치 인덱스 집합을 의미한다. 이 과정은 고립된 노이즈 패치의 영향을 억제하고, 객체가 존재하는 영역의 점수를 보강하여 클러스터 형태의 토큰 보존을 유도한다. 결과적으로 비전 맘바 인코더 블록은 끊기지 않는 연속된 토큰 시퀀스를 입력 받게 되어, 재학습 없이도 안정적인 문맥 추론이 가능해진다.

Ⅱ-3. 원스텝 프루닝 및 패치 재정렬

프루닝을 통해 선별된 핵심 패치들은 비전 맘바 인코더에 입력되기 전, 원래의 스캔 순서대로 재정렬된다. 이는 비전 맘바 패치들의 위치 임베딩과 패치 간의 인접 정보를 보존하기 위함이다. 본 연구는 인코더 진입 전 단 한 번의 프루닝으로 전체 연산량을 절감하는 원스텝 구조를 채택함으로써, 기존 다단계 프루닝 대비 추론 지연 시간을 대폭 단축하였다.

Ⅲ. 실험 및 결과

본 연구에서는 ViM-Small 모델을 사용하여 ImageNet-1K 데이터셋에 대한 분류 정확도를 평가하였다. 제안하는 기법들의 효율성을 검증하기 위해, 모든 실험은 별도의 재학습이 없는 조건에서 수행되었다. 비교군으로는 단순 L2-Norm 기반의 Naïve 프루닝과 최신 비전 맘바 토큰 감소 기법인 EViT[4], PuMer[5], UTR[6], HSA[7]로 설정하였다.

실험 결과, 본 논문에서 제안한 패치 프루닝에 스코어 스무딩 기법을 추가한 방법은 표 1과 같이 기존 기법들 대비 더 높은 정확도를 보여주었다.

표 1. 토큰 프루닝 기법들의 성능 비교 분석

Method	Token Reduction	Params (M)	Top-1 Acc. (%)	Δ
ViM-Small	0%	26	80.5	0.0
+ EViT	20%	26	75.8	4.7↓
+ PuMer		26	76.9	3.6↓
+ UTR		26	77.3	3.2↓
+ HSA		26	76.7	3.8↓
+ Ours (PaPr+SS)		**26**	**78.2**	**2.3↓**
+ EViT	30%	26	71.8	8.7↓
+ PuMer		26	74.6	5.9↓
+ UTR		26	75.0	5.5↓
+ HSA		26	74.8	5.7↓
+ Ours (PaPr+SS)		**26**	**76.1**	**4.4↓**
+ EViT	40%	26	64.8	15.7↓
+ PuMer		26	69.1	11.4↓
+ UTR		26	71.5	9.0↓
+ HSA		26	71.2	9.3↓
+ Ours (PaPr+SS)		**26**	**73.8**	**6.7↓**

토큰을 40% 제거했을 때, 기존 ViT 기반 기법인 EViT(64.8%)나 Mamba 전용 기법인 UTR(71.5%)보다 본 연구에서 제안한 PaPr + SS(73.83%)의 성능이 더 높게 측정되었다. 이는 스코어 스무딩이 ViM 의 시퀀스 모델링에 필수적인 지역적 일관성을 효과적으로 보존함을 의미한다. 단순히 PaPr 방법만 중요도를 예측했을 때(70.94%)보다 Score Smoothing 을 결합했을 때 정확도가 약 2.89%p 추가 향상되는 결과를 보였다. 이는 단일 패치의 중요도뿐만 아니라 주변 패치와의 공간적 맥락을 함께 고려하는 것이 비전 맘바 구조에서 매우 중요한 요소임을 입증한다. 본 논문에서 제안한 기법은 추가적인 파인 튜닝 없이도 70% 이상의 높은 정확도를 유지하며, 이는 자원이 제한된 에지 디바이스에서 재학습 비용 없이 즉각적인 경량화 모델 배포가 가능함을 시사한다.

표 2. 제안한 스코어 스무딩 적용 유무에 따른 정확도 분석

Method	Token Reduction	Top-1 Acc. (%)	Δ
ViM-Small	0%	80.5	0.0
+ PaPr	20%	76.9	3.6↓
+ (PaPr+SS)		**78.2**	**2.3↓**
+ PaPr	30%	74.4	6.1↓
+ (PaPr+SS)		**76.1**	**4.4↓**
+ PaPr	40%	70.9	9.6↓
+ (PaPr+SS)		**73.8**	**6.7↓**
+ PaPr	50%	66.3	14.2↓
+ (PaPr+SS)		**72.4**	**8.1↓**

Ⅳ. 결 론

기존의 토큰 프루닝 기법들은 비전 트랜스포머의 이차적 복잡도($O(N^2)$) 문제를 해결하는 데는 효과적이었으나, 순차적인 데이터 처리를 수행하는 비전 맘바 구조에서는 정보 흐름의 단절을 초래하여 성능 저하를 일으킨다는 한계를 확인하였다. 이를 해결하기 위해 본 논문은 비전 맘바의 순차적 특성을 고려한 스코어 스무딩 기반 패치 프루닝 기법을 제안한다. 경량 CNN 으로 패치 중요도를 예측하고 스무딩 필터로 공간적 맥락을 보정함으로써, 정보 단절 없는 히든 스테이트 갱신을 구현하였다. 실험 결과, 40% 토큰 제거 시 73.8%의 정확도를 기록해 기존 UTR(71.5%) 대비 우수한 성능을 입증하였으며, 재학습 없이도 추론이 가능한 수준의 효율성을 확보하였다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00409492).

참 고 문 헌

[1] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model," in Proc. 41st International Conference on Machine Learning (ICML), 2024.

[2] T. Mahmud et al., "Papr: Training-free one-step patch pruning with lightweight convnets for faster inference," in Proc. European Conference on Computer Vision (ECCV), Cham: Springer Nature Switzerland, 2024, pp. 110-128.

[3] Gonzalez, R. C., & Woods, R. E. (2018). Digital Image Processing. Pearson.

[4] Y. Pan, Z. Liu, H. Liu, J. Ma, S. Ge, and Y. Wang, "EViT: Expediting Vision Transformers via Token Reorganizations," in Proc. International Conference on Learning Representations (ICLR), 2022.

[5] Q. Cao, B. Paranjape, and H. Hajishirzi, "PuMer: Pruning and Merging Tokens for Efficient Vision Language Models," in Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 12890-12903.

[6] Z. Zhan et al., "Rethinking Token Reduction for State Space Models," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 1686-1697.

[7] Z. Zhan et al., "Exploring Token Pruning in Vision State Space Models," in Proc. 38th Conference on Neural Information Processing Systems (NeurIPS), 2024.