레터논문 (Letter Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.181
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

멀티모달 LLM에서의 비전 인코더 계층별 엔트로피 변화를 활용한 시각 토큰 프루닝

백 창 우[a)*], 김 소 현[a)*], 송 주 원[b)*], 백 광 렬[a)], 공 경 보[a)‡]

Visual Token Pruning Based on Entropy Dynamics in Vision Encoders of Multimodal LLMs

Changwoo Baek[a)*], Sohyeon Kim[a)*], Jouwon Song[b)*], Kwang-Ryul Baek[a)], and Kyeongbo Kong[a)‡]

요 약

토큰 프루닝(Token pruning)은 멀티모달 대형 언어 모델(Multimodal Large Language Models, MLLMs)의 효율성을 향상시키는 핵심 기법으로 부상하고 있다. 그러나 대부분의 기존 접근법은 고정된 비율 또는 깊은 층의 어텐션 신호에 의존한다. 본 연구에서는 이러한 한계를 극복하고 프루닝 과정을 체계적으로 이해하기 위해, MLLM 비전 인코더 내에서 깊이에 따른 어텐션의 재분배 양상을 분석하였다. [CLS] 어텐션의 층별 분석 결과, 얕은 층은 높은 엔트로피의 다양한 패턴을 보이고, 깊은 층은 낮은 엔트로피의 전역 요약으로 수렴하며, 중간 층에서는 급격한 엔트로피 감소가 나타나 위상 전이(phase transition)를 형성함을 확인하였다. 이러한 분석을 바탕으로, 우리는 성능과 효율성을 재학습 없이 균형 있게 조절하는 엔트로피 기반 적응적 층 선택(entropy-based adaptive layer selection) 방법을 제안하며, 이는 효율적인 MLLM 설계를 위한 실질적인 지침을 제공한다.

Abstract

Token pruning has emerged as a key technique for improving the efficiency of multimodal large language models (MLLMs), yet most approaches rely on fixed or deep-layer attention signals. To better understand and optimize pruning, we analyze how attention is redistributed across depth within MLLM vision encoders. A layer-wise examination of [CLS] attention reveals a consistent depth-dependent pattern: shallow layers exhibit high-entropy, diverse attention; deep layers converge into low-entropy global summaries; and mid layers experience a sharp entropy drop, marking a phase transition. Building on these findings, we propose an entropy-based adaptive layer selection that balances pruning efficiency and performance without retraining, offering practical guidance for efficient MLLM design.

Keyword : Multimodal Large Language Models, Vision-Language Models, Visual Token Pruning, Inference Optimization, On-device AI

Ⅰ. 서 론

최근 멀티모달 대형 언어 모델(MLLMs)[1]은 뛰어난 성능을 보이지만, 입력 이미지를 수백 개의 시각 토큰으로 처리함에 따라 연산량이 증가하고 추론 속도가 저하되는 한계를 가진다. 이를 완화하기 위해 시각 토큰 프루닝 기법[2][3][4]이 제안되었으나, 기존 방법들은 주로 비전 인코더의 단일 고정 층, 특히 깊은 층의 어텐션에 의존하여 토큰 중요도를 판단한다. 이러한 접근은 중요한 시각 정보가 항상 마지막 계층에 집중된다는 가정에 기반하나, 실제로는 입력에 따라 해당 가정이 성립하지 않는다. 본 연구는 트랜스포머 기반 비전 인코더 내부의 어텐션 분포를 정량적으로 분석하여, 계층 깊이에 따른 엔트로피 변화 패턴을 규명한다. 분석 결과, 얕은 층에서는 어텐션이 넓게 분포하여 높은 엔트로피를 보이는 반면, 깊은 층으로 갈수록 일부 핵심 토큰에 집중되어 낮은 엔트로피를 나타낸다. 또한 중간 계층에서 엔트로피가 급격히 감소하는 전이 구간이 관찰되었으며, 이는 지역적 특징 탐색에서 전역적 의미 통합으로의 전환을 반영한다. 이러한 엔트로피 동역학은 이미지 복잡도에 따라 달라지는 입력 의존적 특성을 보인다. 이러한 관찰을 바탕으로, 본 연구는 비전 인코더를 얕은 층과 깊은 층으로 구분하고, 각 구간에서 대표 계층을 입력별로 적응적으로 선택하는 층 선택 기반 시각 토큰 프루닝 기법을 제안한다.

Ⅱ. 실증적 분석

본 절에서는 멀티모달 대형 언어 모델의 비전 인코더 내부에서 어텐션이 깊이에 따라 어떻게 재분배되는지를 실증적으로 분석하였다. 이를 위하여 CLIP 기반 비전 인코더를 대상으로 각 계층 l에서 [CLS] 토큰이 패치 토큰에 부여하는 어텐션 분포 $S^l \in R^N$을 추출하였다. 여기서 N은 패치 토큰의 개수를 의미한다. 각 계층의 정보 집중도를 정량화하기 위해 Shannon 엔트로피 $H(l)$을 다음과 같이 정의하였다.

$$H(l) = -\sum_{j=1}^{N} S^l \log S_j^l \tag{1}$$

실험 결과, 그림 1에서 보이듯 어텐션 엔트로피는 레이어가 깊어질수록 전반적으로 감소하는 경향을 보였다. 이는 얕은 층에서는 어텐션이 여러 토큰에 넓게 분산되는 반면, 깊은 층에서는 소수의 중요한 토큰에 집중되기 때문이다. 즉, 비전 인코더는 초기에는 다양한 지역적 정보를 탐색하고, 후반으로 갈수록 이를 압축하여 전역적인 요약을 형성한다. 그림 2의 어텐션 시각화 역시 얕은 층의 분산된 분포와 깊은 층의 집중된 분포를 보여주며 이러한 엔트로피 감소 경향을 뒷받침한다.

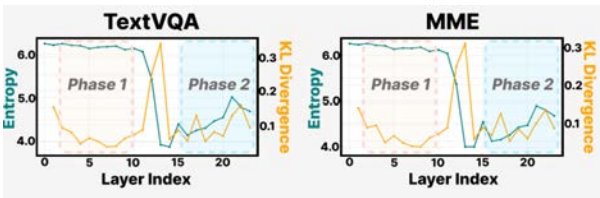

그림 1. 중간 레이어에서의 엔트로피 및 KL 발산의 전이 현상
Fig. 1. Entropy and KL-divergence transition at mid layers

또한 엔트로피는 단조 감소가 아니라, 중간 레이어 부근에서 급격히 감소하는 구간이 존재함을 확인하였다. 이러한 변화의 크기를 정량화하기 위해, 인접한 레이어 간 어텐션 분포 차이를 KL-divergence로 측정하였다. 그 결과,

$$D_{KL}(S^j \parallel S^{j+1}) = \sum_{j=1}^{N} S^l \log \frac{S_j^l}{S_j^{l+1}} \tag{2}$$

중간 레이어 구간에서 D_{KL} 값이 갑자기 크게 증가하는 스파이크(spike)가 나타났으며, 이는 토큰 중요도가 급격히 재정렬되는 구간임을 의미한다. 즉, 모델은 얕은 층에서 다양한

a) 부산대학교 전기전자공학부(Department of Electrical & Electronics Engineering, Pusan National University)
b) LG전자(LG Electronics)
* Equal Contribution
‡ Corresponding Author : 공경보(Kyeongbo Kong)
E-mail: kbkong@pusan.ac.kr
Tel: +82-51-510-2399
ORCID: https://orcid.org/0000-0002-1135-7502
※ This work was supported by a 2-Year Research Grant of Pusan National University.
· Manuscript October 24, 2025; Revised December 4, 2025; Accepted December 4, 2025.

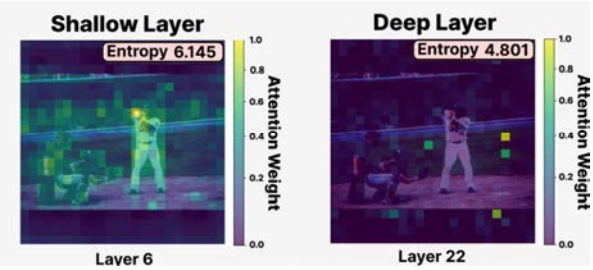

그림 2. 레이어별 확산형 및 응집형 어텐션 패턴
Fig. 2. Dispersed and concentrated attention patterns in shallow and deep layers

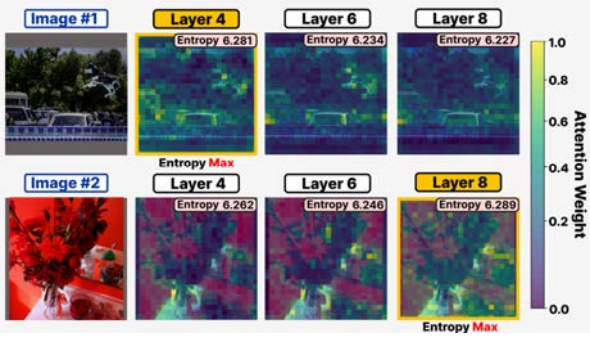

그림 3. 입력 이미지별 최대 엔트로피 레이어의 상이성
Fig. 3. Variation of maximum-entropy layers across input images

토큰을 탐색한 뒤 중간 층에서 핵심 정보를 선별하고, 깊은 층에서 압축된 정보에 집중한다. 또한 그림 3에서 보이듯 엔트로피의 최대·최소 지점은 이미지마다 달라, 엔트로피 변화가 입력 이미지에 따라 달라지는 동적 특성임을 확인할 수 있다.

Ⅲ. 제안하는 방법

실증적 분석을 통해 층별 엔트로피 분기 현상과 입력 의존적 계층 동역학을 확인하였다. 얕은 층은 높은 엔트로피로 지역적 다양성을 유지하고, 깊은 층은 낮은 엔트로피로 전역적 의미를 응축한다. 또한 최대·최소 엔트로피를 갖는 계층은 입력 특성에 따라 달라지므로, 고정된 층을 사용하는 방식보다 입력별로 동적으로 계층을 선택하는 접근이 효과적이다. 이에 따라 본 연구는 각 입력 이미지에 대해 엔트로피가 최대인 얕은 층과 최소인 깊은 층을 선택하며, 선택 과정은 식 (3)과 같이 정의된다.

$$l_s = \arg\max_{l \in L_{shallow}} H(l),\ l_d = \arg\min_{l \in L_{deep}} H(l) \quad (3)$$

얕은 층 l_s은 지역 정보의 다양성을, 깊은 층 l_d은 전역적 의미를 담당하므로, 두 층에서 각각 전체 토큰의 일부를 선택하되 그 비율의 합은 1로 제한한다. 선택된 토큰을 결합하고, 제거된 토큰은 가장 유사한 유지 토큰에 병합하여 프루닝 이후에도 시각적 다양성과 의미적 일관성을 보존한다.

Ⅳ. 실험

1. 주요 실험 결과

본 연구에서는 제안된 방법을 널리 사용되는 오픈소스 모델 LLaVA-1.5-7B에 적용하였다. 비전 토큰의 선택 비율은 각각 p_s= 0.4, p_d = 0.6으로 설정하였다. 6 - 8번째 레이어는 지역적 특징을 포착하기 위해 선택되었으며, 20 - 21번째 레이어는 심층적 의미 정보를 표현하기 위해 선택되었다.

표 1. LLaVA-1.5-7B 모델이 64개의 시각 토큰을 유지할 때의 9개 멀티모달 벤치마크 결과
Table 1. Results of LLaVA-1.5-7B on nine multimodal benchmarks when retaining 64 visual tokens

Method	TextVQA[4]	MME[5]	VizWiz[8]	Avg.
LLaVA-1.5-7B	58.2	1506	50.1	100.0%
FastV	51.6	973	49.1	83.7%
SparseVLM	52.1	1190	49.4	89.0%
VisionZip	55.7	1365	52.9	97.3%
DivPrune	54.5	1334	53.6	96.4%
Ours	56.0	1416	54.1	99.4%

표 2. LLaVA-Next-7B 모델이 320개 시각 토큰을 유지할 때의 9개 멀티모달 벤치마크 결과
Table 2. Results of LLaVA-Next-7B on nine multimodal benchmarks when retaining 320 visual tokens

Method	TextVQA[4]	MME[5]	VizWiz[8]	Avg.
LLaVA-Next-7B	60.3	1512	55.2	100.0%
FastV	52.2	1099	51.3	84.1%
SparseVLM	56.5	1386	54.2	94.5%
VisionZip	58.8	1444	56.2	98.2%
DivPrune	56.2	1423	55.6	96.0%
Ours	58.5	1465	55.8	98.3%

표 1에서 보이듯이, 원본 576개의 비전 토큰을 64개(약

11%)로 축소하는 공격적인 프루닝 설정에서도, 3개 벤치마크 평균 성능 저하는 0.6%에 불과하였다. 이는 FastV[2], SparseVLM[3]와 같은 LLM 어텐션 기반 방법뿐 아니라 VisionZip[6] 및 유사도 기반 방법인 DivPrune[7]보다도 우수한 성능을 기록하였다. 더 나아가, 표 2와 같이 2,880개의 비전 토큰을 사용하는 LLaVA-NeXT-7B에서도 320개 토큰만 유지한 설정에서다른 베이스라인 대비 가장 높은 성능을 달성하였다.

표 3에서는 레이어 선택 전략에 따른 성능을 비교하였다. 고정 레이어 쌍(L6 – L20)은 기준 성능을 보였고, adaptive 전략 중 Min – Max와 Min – Min은 성능이 낮았다. Max – Max는 일부 개선을 보였으나, 제안한 Max – Min 전략이 가장 우수했으며, 이는 얕은 층의 다양성과 깊은 층의 전역 정보를 조합하는 것이 효과적임을 보여준다.

표 3. 엔트로피 기반 레이어 선택에 대한 ablation 실험
Table 3. Ablation study of entropy-based layer selection

Strategy	TextVQA	MME
L6 – L20	55.70	1389
Min – Max	55.62	1362
Min – Min	55.64	1379
Max – Max	55.84	1401
Max – Min (Ours)	56.02	1416

표 4. 주요 연산 지표 비교
Table 4. Comparison of Key Computational Metrics

Method	#Tokens	FLOPs ↓	Accuracy ↑
Default	576	3.82T	58.2
FastV	64	1.07T	51.6
SparseVLM	64	1.08T	52.1
Ours	64	0.42T	56.0

2. 효율성 분석

표 4에 나타난 바와 같이, 제안된 방법은 TextVQA 데이터셋에서 64개 토큰만을 사용하여 프루닝을 수행했음에도, 기존 LLaVA-1.5-7B 모델 대비 96.2%의 원본 성능을 유지하였다. FastV 및 SparseVLM 등 기존 방법들은 LLM 내부에서 토큰을 선택 및 제거하지만, 우리의 방법은 LLM 입력 이전 단계에서 프루닝을 수행하여 LLM 내부 연산량을 크게 감소시킨다.

V. 결 론

본 연구에서는 비전 인코더의 계층별 엔트로피 변화를 분석하여, 입력별로 적응적으로 층을 선택하는 새로운 시각 토큰 프루닝 기법을 제안하였다. 제안된 방법은 기존 방식 대비 높은 효율성과 성능 유지율을 동시에 달성하였으며, 멀티모달 모델의 추론 효율화를 위한 효과적인 방향을 제시한다.

참 고 문 헌 (References)

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” Advances in Neural Information Processing Systems, Vol.36, pp.34892 – 34916, 2023.
doi: https://doi.org/10.48550/arXiv.2304.08485

[2] L. Chen, H. Zhao, T. Liu, S. Bai, J. Lin, C. Zhou, and B. Chang, “An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models,” Proceedings of the European Conference on Computer Vision (ECCV), pp.19 – 35, Springer, 2024.
doi: https://doi.org/10.1007/978-3-031-73004-7_2

[3] Y. Zhang, C. Fan, J. Ma, W. Zheng, T. Huang, K. Cheng, D. Gudovskiy, T. Okuno, Y. Nakata, K. Keutzer, et al., “SparseVLM: Visual token sparsification for efficient vision-language model inference,” Proceedings of the International Conference on Machine Learning (ICML), 2025.
doi: https://doi.org/10.48550/arXiv.2410.04417

[4] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, “Towards VQA models that can read,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.8317 – 8326, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00851

[5] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, et al., “MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models,” Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2025.
doi: https://doi.org/10.48550/arXiv.2306.13394

[6] S. Yang, et al., “VisionZip: Longer is better but not necessary in vision-language models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
doi: https://doi.org/10.1109/CVPR52734.2025.01843

[7] S. R. Alvar, et al., “DivPrune: Diversity-based visual token pruning for large multimodal models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
doi: https://doi.org/10.1109/CVPR52734.2025.00877

[8] J. P. Bigham, et al., “VizWiz: Nearly real-time answers to visual questions,” Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST), pp.333 – 342, 2010.
doi: https://doi.org/10.1145/1866029.1866080