2025년도 한국통신학회 하계종합학술발표회

에너지 효율성 증진을 위한 Transformer 최적화 기술 최신 연구 동향

정재현, 안효준, 김규선, 김중헌
고려대학교
{rupang1234, hyojun, kingdom0545, joongheon}@korea.ac.kr

Recent Trends in Transformer Optimization for Improved Energy Efficiency

Jaehyun Chung, Hyojun Ahn, Gyu Seon Kim, Joongheon Kim
Korea Univ.

요 약

최근 인공지능 기술의 발전과 함께 Transformer 기반 모델이 다양한 응용 분야에서 높은 성능을 보이며 주목받고 있다. 그러나 모델의 대규모화로 인해 연산량과 메모리 소비가 증가함에 따라 에너지 효율성에 대한 문제가 제기되고 있다. 본 논문에서는 에너지 효율성을 향상시키기 위한 Transformer 최적화 기술들을 조사한다. 또한, 다양한 최적화 기법들의 구조적 특징과 적용 사례를 분석하여 경량화된 Transformer 모델의 가능성과 한계를 탐색한다.

Ⅰ. 서 론

딥러닝 기술의 급속한 발전과 함께 Transformer 모델은 자연어 처리, 컴퓨터 비전, 음성 인식 등 다양한 분야에서 기존 모델을 뛰어넘는 성능을 보여주며 핵심 기술로 자리잡고 있다. 특히, 대규모 사전학습을 통해 범용성과 전이 학습 효율성을 동시에 확보할 수 있다는 장점으로 인해 산업 전반에 걸쳐 폭넓게 활용되고 있다. 그러나 이러한 성능 향상은 막대한 계산 자원과 에너지 소비를 수반하며, 이는 실시간 시스템, 엣지 디바이스, 탄소 배출 감소 등과 같은 현실적 제약에 큰 부담으로 작용한다. 최근에는 이와 같은 문제를 해결하기 위해 Transformer 모델의 경량화와 연산 최적화를 목표로 하는 다양한 연구가 진행되고 있으며, 이는 에너지 효율성과 환경 지속 가능성 측면에서도 중요한 기술적 과제로 인식되고 있다. 특히, 연산량 감소, 파라미터 수 축소, 동적 추론 구조 도입, 저정밀 양자화 등의 기법은 성능 저하 없이 에너지 소비를 줄일 수 있는 방향으로 주목받고 있다. 이러한 기술들은 고성능 AI를 제한된 자원에서 실행해야 하는 모바일 디바이스 및 IoT 환경에서 높은 실용성을 가진다. 본 논문에서는 이러한 최적화 기법들의 기술적 흐름과 구조적 특징을 분석하고, 에너지 효율성 증진이라는 관점에서 Transformer의 미래 지향적 설계 방안을 모색하고자 한다.

Ⅱ. 본론

2-1 Transformer 구조와 에너지 효율성 과제

Transformer는 [그림 1]에서와 같이] Attention 메커니즘을 중심으로 설계된 딥러닝 모델로, 특히 자연어 처리 분야에서 획기적인 성능을 보여주며 기존 RNN 및 CNN 기반 모델을 빠르게 대체하였다 [1]. 이후 Vision Transformer (ViT), Swin Transformer와 같은 확장 모델이 등장하면서 영상 처리와 다중 모달 학습 등 다양한 영역으로 적용 범위를 넓혀왔다. Transformer의 핵심은 병렬 처리가 가능한 Self-Attention 구조이며, 이는 Sequence 전반의 의존 관계를 효율적으로 학습하는 데 효과적이다. 하지만 이러한 구조는 입력 길이에 따라 연산량이 제곱으로 증가하는 구조적 한계를 지니며, 이는 실제 배포 시 에너지 소모 및 지연 시간 증가로 이어진다.

특히, 수억 개 이상의 파라미터를 갖는 대규모 사전학습 모델은 GPU 자원을 과도하게 사용하며, 엣지 디바이스나 배터리 기반 환경에서는 실용성이 떨어지는 문제가 있다. 예를 들어, GPT-3는 단일 질의 응답에 수십~수백 Wh의 에너지를 소모할 수 있으며, 이는 대규모 사용자 기반 서비스에 큰 비용 부담을 초래한다. 따라서 Transformer 기반 모델의 에너지 효율성 증진은 연구 커뮤니티와 산업계에서 중요한 과제로 부상하고 있다. 최근에는 연산 최적화와 구조 경량화를 중심으로 다양한 접근이 제안되고 있으며, 이는 환경적 지속가능성과 경제성 측면에서도 핵심적인 기술 개발 방향으로 간주된다.

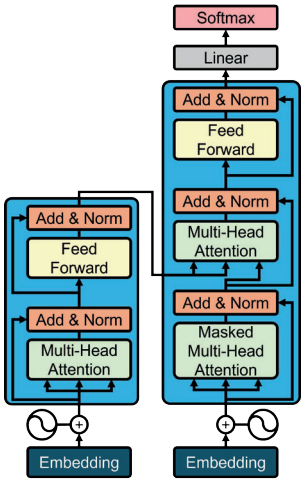

그림 1. Transformer 구조

2-2 에너지 효율적 Transformer 최적화 기술 관련 연구

Transformer 모델의 높은 연산 비용과 메모리 요구는 저전력 환경이나 실시간 응용에서 실용적인 제약 요인으로 작용해 왔으며, 이를 해결하기 위한 다양한 최적화 연구가 활발히 이루어지고 있다. 한 연구에서는 Transformer의 다중 어텐션 구조에 주목하여, 모든 어텐션 헤드가 항상 동일한 중요도를 가지지 않는다는 점을 실험적으로 보였다 [2]. 실험 결과,

중요도가 낮은 일부 헤드를 제거해도 모델 성능 저하가 거의 발생하지 않았으며, 이를 통해 연산량을 효과적으로 줄이고 에너지 소비를 절감할 수 있는 가능성을 제시하였다. 이러한 접근은 모델 경량화뿐 아니라 해석 가능성 개선에도 기여할 수 있다. 또 다른 연구에서는 Transformer를 하드웨어에서 효율적으로 구현하기 위한 구조 변환 기법을 제안하였다 [3]. 해당 연구에서는 모델의 행렬 연산을 Block-Circulant 형태로 구조화하고, 이를 FPGA 환경에서 최적화함으로써 모델의 파라미터 수를 최대 16배까지 줄였다. Block-Circulant 행렬의 순환적 특성을 활용하여 연산 효율을 극대화하였으며, 실험을 통해 기존 GPU 대비 약 8배, CPU 대비 27배 이상의 에너지 효율 향상 효과가 확인되었다. 이는 Transformer 기반 모델을 엣지 디바이스에서도 실용적으로 운영할 수 있게 만드는 중요한 기반이 된다. Transformer의 추론 단계를 최적화하려는 시도도 이루어졌다. 한 연구에서는 시계열 분류 문제를 대상으로 구조적 프루닝과 정량화 기법을 병행 적용함으로써, 연산량을 줄이면서도 정확도를 유지하는 방법을 제안하였다 [4]. 정량화는 부동소수점 연산을 정수 기반 연산으로 치환하여 연산 속도를 향상시키고 전력 소비를 줄였으며, L1 기반 프루닝은 중요도가 낮은 연결을 제거해 모델 구조를 간결화하였다. 실험 결과, 추론 속도는 60% 이상 향상되었고 에너지 소비는 약 30% 감소하였다.

Self-Attention 구조의 연산 효율성을 개선하기 위한 연구는 크게 두 가지 방향으로 전개되었다. 하나는 스파이킹 뉴럴 네트워크(Spiking Neural Network, SNN)의 이벤트 기반 연산 구조를 Transformer에 적용하여 연산 에너지를 대폭 절감하는 접근이며, 다른 하나는 Self-Attention의 연산 복잡도를 줄이기 위한 Sparse Attention 구조 설계이다. 전자의 경우, 곱셈 중심의 연산을 마스크 기반 덧셈 연산으로 대체하거나 스파이크 이벤트 기반 계산 흐름을 도입하여 기존 대비 최대 87배에 달하는 연산 에너지 절감을 보고하였다 [5]. 후자의 경우, Self-Attention을 Quadratic에서 Linear 혹은 Sublinear 복잡도로 근사하는 다양한 구조가 제안되었으며, 이를 Edge-Optimized Transformer에 통합함으로써 Sequence 길이가 긴 입력에서도 전력 소모와 처리 속도 모두를 효율적으로 개선하였다 [6]. 이러한 방법들은 공통적으로 입력 데이터의 특성과 중요도에 따라 연산을 선택적으로 수행하는 구조를 도입하여, 모바일 디바이스, 자율주행 센서, IoT 환경 등에서 에너지 효율적 추론을 가능하게 한다. 또 다른 연구에서는 Transformer 최적화를 위한 알고리즘적 기법과 하드웨어 설계를 통합적으로 고려하였다 [6]. 이 연구에서는 Quantization-Aware Training(QAT), Structured Pruning, Dynamic Sparsity 등과 같은 기법이 실제 칩 설계에서 어떻게 활용될 수 있는지를 분석하고, 메모리 접근 최소화, 병렬 연산 블록 구성, 연산 스케줄링 등 시스템 수준에서의 구현 전략을 함께 제시하였다. 이를 통해 Transformer 모델을 FPGA나 ASIC 환경에 효과적으로 배치할 수 있는 설계 기준을 제공하였다. 마지막으로, 동적 추론 기법을 통해 상황에 따라 연산 경로를 조정하고, 불필요한 연산을 실시간으로 생략하는 접근도 제안되었다 [7]. 이 방식은 입력 난이도나 중요도에 따라 활성화 블록을 선택적으로 조정하며, 평균 연산량과 전력 소비를 대폭 줄이는 동시에 정확도를 유지하는 것이 특징이다. Sparsity-Aware Accelerator를 활용하여 데이터 재사용을 극대화하고, 메모리 대역폭과 계산 자원을 함께 절약함으로써 에너지 효율을 극대화하였다.

Ⅲ. 결론

본 논문은 Transformer 모델의 에너지 효율성을 향상시키기 위한 최적화 기술들을 구조적, 알고리즘적, 하드웨어적 측면에서 고찰하였다. Self-Attention 연산의 복잡도를 줄이는 구조 개선, 파라미터 수 축소를 위한 프루닝과 정량화, 스파이킹 기반의 이벤트 중심 연산, 그리고 입력 중요도에 따른 동적 추론 기법 등은 공통적으로 성능 저하 없이 연산 자원과 전력 소비를 절감하는 데 효과적임을 보였다. 이러한 기술들은 특히 엣지 디바이스, 모바일 환경, IoT 시스템 등 에너지 제약이 큰 응용 분야에서 Transformer 모델의 실용성을 높이는 데 중요한 역할을 한다. 더불어 하드웨어-소프트웨어 통합 최적화를 통해 실제 시스템 수준에서의 효율적인 구현 가능성도 함께 제시되고 있다. 향후에는 다양한 기법들의 결합 효과에 대한 체계적 분석과, 에너지 소비에 따라 연산을 적응적으로 조절할 수 있는 모델 설계가 주요 연구 과제로 부상할 것으로 기대된다. Transformer의 에너지 효율성 개선은 지속 가능한 고성능 AI 시스템 구현을 위한 핵심 기술로, 앞으로도 활발한 연구와 기술 발전이 요구된다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임
(IITP-2025-RS-2024-00436887).

참 고 문 헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, December 2017, pp. 5998 - 6008.

[2] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, December 2019, pp. 14014 - 14024.

[3] B. Li, S. Pandey, H. Fang, Y. Lyv, J. Li, J. Chen, M. Xie, L. Wan, H. Liu, and C. Ding, "FTRANS: Energy-efficient acceleration of transformers using FPGA," in *Proc. ACM/IEEE Int. Symp. on Low Power Electronics and Design (ISLPED)*, Boston, MA, USA, August 2020, pp. 175 - 180.

[4] A. Kermani, E. Zeraatkar, and H. Irani, "Energy-efficient transformer inference: Optimization strategies for time series classification," *CoRR*, vol. abs/2502.16627, February 2025.

[5] C. Du, Q. Wen, Z. Wei, and H. Zhang, "Energy efficient spike transformer accelerator at the edge," *Intelligent Marine Technology and Systems*, vol. 2, no. 1, p. 24, September 2024.

[6] M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, and G. Li, "Spike-driven transformer," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, December 2023, pp. 64043-64058.

[7] S. Huang, E. Tang, S. Li, X. Ping, and R. Chen, "Hardware-friendly compression and hardware acceleration for transformer: A survey," *Electronic Research Archive*, vol. 30, no. 10, pp. 3755-3785, August 2022.

[8] S. Tuli and N. K. Jha, "AccelTran: A sparsity-aware accelerator for dynamic inference with transformers," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 11, pp. 4038 - 4051, November 2023.