탄소중립 실현을 위한 대규모 언어모델 경량화 기반 에너지 최적화 연구 동향

안효준, 박수현*, 김중헌
고려대학교, *숙명여자대학교

hyojun@korea.ac.kr, *soohyun.park@sookmyung.ac.kr, joongheon@korea.ac.kr

Research Trends in Energy Optimization Based on Large Language Model Compression for Achieving Carbon Neutrality

Hyojun Ahn, Soohyun Park*, Joongheon Kim
Korea Univ., *Sookmyung Women Univ.

요 약

본 논문은 대규모 언어 모델(LLMs)의 학습 및 추론 과정에서 발생하는 에너지 소비와 탄소 배출 문제를 분석하고, 이를 해결하기 위한 모델 경량화 및 에너지 최적화 기술 동향을 고찰한다. 인공지능 분야의 급속한 발전과 함께 LLM 규모가 기하급수적으로 증가함에 따라 모델 학습에 필요한 컴퓨팅 자원과 에너지 소비량도 크게 증가하고 있다. 본 연구에서는 양자화, 지식 증류, Pruning, 혼합 전문가 모델과 같은 경량화 기술과 탄소 인식 인프라 활용, 하드웨어 최적화, 적응형 정밀도 학습법 등 에너지 효율을 높이기 위한 다양한 접근법을 살펴본다. 또한 LLM의 탄소 발자국 측정 및 평가 방법론에 대해 논의하고, 탄소중립 실현을 위한 향후 연구 방향을 제시한다. 궁극적으로 AI 환경적 영향을 최소화하는 노력은 탄소중립 목표 달성에 기여한다.

Ⅰ. 서 론

최근 GPT-4, Claude, LLaMA 와 같은 대규모 언어 모델(Large Language Models, LLMs)은 자연어 처리 분야에서 혁신적인 성능을 보이며 다양한 산업 분야에 적용되고 있다. 그러나 이러한 LLM 의 학습과 추론 과정은 막대한 컴퓨팅 자원을 필요로 하며, 이는 상당한 에너지 소비와 탄소 배출로 이어진다. GPT-3 175B 모델의 학습 과정에서는 약 1,287 MWh 의 전력이 소비되었으며, 이는 약 552 톤의 CO_2 배출량에 해당한다 [1]. 디지털 전환이 가속화되는 가운데, AI 모델의 에너지 효율성 향상은 탄소중립 실현을 위한 핵심 과제로 부상하고 있다. 본 논문에서는 LLM 의 경량화 기술과 에너지 최적화 방법론에 대한 최신 연구 동향을 살펴보고, 향후 지속 가능한 AI 발전 방향을 제시하고자 한다.

Ⅱ. LLM의 에너지 소비 현황과 환경적 영향

그림 1 과 같이 LLM 의 규모가 커짐에 따라 학습 및 추론에 필요한 에너지 소비량은 기하급수적으로 증가하고 있다. 대규모 Transformer 기반 모델을 학습시키는 과정에서 발생하는 탄소 배출량은 자동차 5 대가 평생 배출하는 양과 맞먹는다 [2]. 특히 모델 크기와 연산량이 증가할수록 이러한 환경적 영향은 더욱 심화된다. AI 모델의 추론 과정에서도 상당한 에너지가 소비되며, 전 세계적으로 배포된 AI 시스템의 누적 탄소 발자국이 중소 국가의 연간 배출량에 맞먹는 수준에 이를 수 있음을 경고하는 연구도 있다. LLM 의 에너지 소비는 학습 단계에서 가장 크게 나타나지만, 수십억 사용자가 매일 추론 과정을 활용하는 환경에서는 추론 단계의 누적 에너지 소비량도 무시할 수 없는 수준에 도달한다. 이는 클라우드 인프라의 확장과 에너지 공급망에도 상당한 부담을 주고 있으며, 데이터센터의 냉각 시스템에 필요한 추가 에너지 소비까지 고려하면 실제 환경 영향은 더욱 클 것으로 예상된다.

Ⅲ. LLM 경량화 기술의 발전

LLM 의 환경적 영향을 줄이기 위한 핵심 접근법으로 모델 경량화 기술이 활발히 연구되고 있다. 양자화(Quantization)는 기존 32 비트 또는 16 비트 부동 소수점 정밀도를 8 비트 또는 4 비트로 낮추는 기법으로, GPT-3 크기의 모델을 4 비트로 양자화했을 때 성능 저하는 5% 미만인 반면 메모리 사용량과 에너지 소비는 최대 8 배까지 감소한다. 지식 증류(Knowledge Distillation)는 대규모 교사 모델의 지식을 작은 학생 모델로 전달하는 기법으로, DistilBERT 는 원본 BERT 모델 대비 파라미터 수를 40% 줄이면서도 97%의 성능을 유지할 수 있다. 가지치기(Pruning)는 모델 내 중요도가 낮은 연결이나 뉴런을 제거하는 기법으로, SparseGPT 방법론은 모델 정확도를 유지하면서도 파라미터의 최대 80%를 제거할 수 있다. 혼합 전문가 모델(Mixture of Experts, MoE)은 계산 효율성이 높은 전문 모듈들을 조합하여 전체 모델을 구성하는 방식으로, Switch Transformer 는 동일한 연산 비용으로 기존 모델보다 더 높은 성능을 달성했다. 이러한 경량화 기술들은 서로 상호 보완적으로 적용될 수 있으며, 최근에는 여러 기법을 조합한 하이브리드 접근법도 연구되고 있다. 특히 양자화와 가지치기를 동시에 적용하는 방법은 단일 기법 적용 대비 더 높은 효율성 향상을 보이는 것으로 나타났다 [3].

IV. 에너지 효율적 학습 및 추론 최적화

모델 구조 자체의 경량화 외에도, 학습 및 추론 과정에서의 에너지 효율을 높이기 위한 다양한 접근법이 연구되고 있다. 학습 최적화 측면에서는 학습 초기에 낮은 정밀도를 사용하고 점진적으로 높이는 적응형 정밀도 학습법을 통해 학습 에너지 소비를 최대 60% 절감할 수 있다. 탄소 인식 인프라 활용 측면에서는 재생 에너지 가용성이 높은 시간대와 지역에 분산 학습을 스케줄링하는 방법을 통해 탄소 배출량을 최대 30% 감소시킬 수 있다 [3]. 하드웨어 최적화 측면에서는 에너지 효율적인 전용 하드웨어 가속기 개발도 활발히 이루어지고 있으며, 저전력 ASIC 을 활용한 추론 최적화를 통해 GPU 대비 에너지 효율을 10 배 이상 개선할 수 있다. 멀티태스킹 및 전이학습 활용 측면에서는 새로운 모델을 처음부터 학습시키는 대신, 기존 모델을 효율적으로 미세 조정하거나 여러 태스크를 동시에 처리하도록 설계함으로써 전체 에너지 소비를 줄이는 방법도 주목받고 있다. 또한 동적 배치 크기 조정, 조기 종료 기법, 학습률 스케줄링 최적화 등의 기법을 통해 불필요한 연산을 줄이고 수렴 속도를 높이는 방법도 에너지 효율성 향상에 기여한다. 클라우드 환경에서는 컴퓨팅 자원의 탄력적 할당과 자동 스케일링을 통해 사용률을 최적화하고 유휴 자원을 최소화하는 접근법도 중요하다 [3].

V. 탄소 발자국 측정 및 평가 기준

LLM 의 탄소 발자국을 정확히 측정하고 평가하기 위한 방법론과 표준화된 지표 개발도 중요한 연구 분야이다. ML 모델의 탄소 배출량을 추적하고 보고하기 위한 "ML CO_2 Calculator"가 개발되었으며, 다양한 AI 시스템의 환경적 영향을 비교 평가할 수 있는 표준화된 프레임워크도 제안되었다. 이러한 측정 도구들은 모델 개발자와 서비스 제공자들이 자신의 AI 시스템이 환경에 미치는 영향을 정량적으로 평가하고, 이를 개선하기 위한 결정을 내리는 데 중요한 역할을 한다. 표준화된 측정 방법론은 또한 다양한 AI 시스템 간의 환경 영향을 공정하게 비교할 수 있게 해주며, 산업 전반의 지속 가능한 AI 개발 문화를 촉진한다 [4]. 최근에는 탄소 발자국 계산에 모델 생애 주기 전체 – 데이터 수집 및 처리, 모델 학습, 배포, 추론, 유지보수, 폐기를 포함하는 포괄적 접근법이 제안되고 있다. 특히 하드웨어 제조 과정에서 발생하는 내재 탄소(embodied carbon)까지 고려한 탄소 회계(carbon accounting) 방법론이 주목받고 있다 [4].

VI. 결론

LLM 의 급속한 성장과 함께 이들의 에너지 소비 및 탄소 배출 문제는 AI 의 지속 가능한 발전을 위한 중요한 과제로 부상했다. 본 논문에서 살펴본 경량화 기술과 에너지 최적화 방법론은 LLM 의 환경적 영향을 최소화하면서도 성능을 유지할 수 있는 가능성을 보여준다. 특히 양자화, 지식 증류, 가지치기와 같은 경량화 기술은 모델 크기를 대폭 줄이면서도 성능 저하를 최소화할 수 있음이 입증되었다. 또한 탄소 인식 인프라 활용과 하드웨어 최적화는 모델 자체의 개선을 넘어 전체 AI 생태계의 지속 가능성을 높이는 데 기여할 수 있다. 향후 연구에서는 더욱 효율적인 모델 구조와 알고리즘 개발뿐만 아니라, 환경적 영향을 고려한 AI 개발 및 배포 가이드라인 수립, 그리고 국제적 협력을 통한 표준화된 측정 방법론 개발이 필요하다. 특히

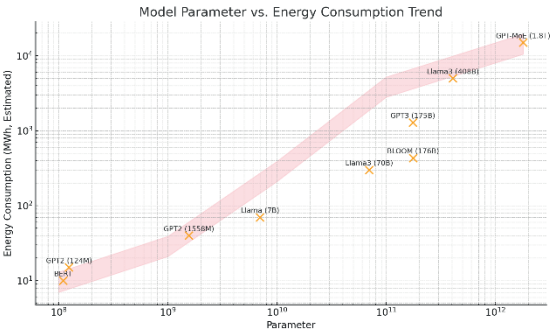

그림 1 모델에 따른 에너지 소비 증가

산업계와 학계의 협력을 통해 발전된 경량화 기술이 실제 배포 환경에서도 효과적으로 적용될 수 있도록 하는 연구가 요구된다. 환경 영향을 최소화하기 위한 모델 설계에서는 성능과 효율성의 균형을 신중하게 고려해야 하며, 특히 지식 증류와 가지치기 기법을 조합한 하이브리드 접근법이 유망하다. 탄소 인식 학습 스케줄링과 에너지 효율적 추론 최적화는 실시간 서비스에서 특히 중요한 역할을 할 것이다. 궁극적으로 AI 의 환경적 영향을 최소화하는 노력은 탄소중립 목표 달성에 기여함과 동시에, 더 넓은 사용자 층이 AI 기술의 혜택을 누릴 수 있는 기회를 확대할 것이다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat, “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model,” *Journal of Machine Learning Research,* vol. 24, no. 1, pp. 11990-12004, Jan., 2023.

[2] Zhongwei *et al.,* “Efficient Large Language Models: A Survey,” *Transactions on Machine Learning Research (TMLR)*, May, 2024.

[3] Lang *et al.*, “A Comprehensive Study on Quantization Techniques for Large Language Models,” in *Proc. International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, Xiamen, China, Dec,, 2024, pp. 224-231

[4] Henderson *et al*, “Towards the systematic reporting of the energy and carbon footprints of machine learning,” *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 10039 – 10081, Jan., 2020