경량화 관점의 소규모 언어 모델 연구동향

임양섭, 정보통신기획평가원, yslim@iitp.kr

Trends on small large langsuage models in quantization

Yangsup Lim, IITP(Institute for Information & Communication Technology Planning & Evaluation)

요 약

최근 인공지능 분야에서 대규모 언어 모델(LLM)의 발전이 주목받고 있다. LLM은 방대한 양의 텍스트 데이터를 기반으로 학습되어 인간 수준의 언어 이해 및 생성 능력을 보여준다. 하지만 LLM은 학습 및 운영에 막대한 양의 데이터와 컴퓨팅 자원을 필요로 하여 실제 서비스에 적용하기 어려운 단점이 있다. 이러한 문제를 해결하기 위해 등장한 소규모 언어 모델(sLLM)은 LLM의 핵심 성능을 유지하면서 매개 변수 수를 줄여 경량화한 모델이다. sLLM은 LLM에 비해 훨씬 적은 데이터와 컴퓨팅 자원으로 학습 및 운영될 수 있어 모바일 기기와 같은 제한된 환경에서도 활용될 수 있다는 장점이 있다. 본 논문에서는 sLLM의 기술적 및 산업적 동향을 분석하고, 온디바이스 AI 제품에 적용된 주요 사례를 살펴봅니다.

Keywords : Small Large Language Model, Quantization, Efficiency

1. 서 론

2020년 6월에 GPT-3가 등장한 이후로 빅테크 주요 기업은 언어 모델(Language Model)의 초거대화 경쟁에 나서고 있다. 빅테크 주요기업이 공개한 LLM을 시간순으로 나열한 것이다[1]. 대규모 언어 모델의 확산과 함께, 최근에는 거대한 클라우드 컴퓨팅 기반의 방대한 데이터 학습이 필요없고, 모바일·PC 등에 탑재하여 별도 기기에서 구현 가능한 소규모 언어 모델(sLLM)이 시장의 주요 트렌드이다. sLLM은 유지 비용이 상대적으로 적게 들고 보안에 대한 걱정도 크지 않으며, 적은 데이터로도 높은 성능을 보여준다. 온디바이스 AI가 보편화되기 위해서는 오프라인 상태에서도 실행할 수 있는, 컴퓨팅 리소스가 작은 언어모델이 필요한 이유이다.

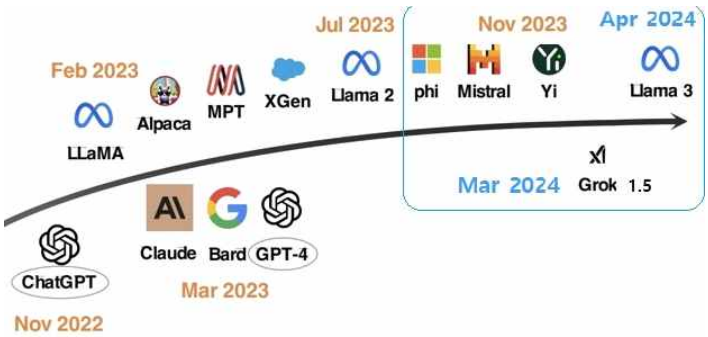

그림 1. LLM 개발 타임라인

2. sLLM 개발 동향

경량화 관점의 소규모 언어 모델 연구동향은 대규모 언어모델의 성능을 유지하면서도 모델의 크기와 계산 복잡도를 줄이려는 노력이다. GPT-4의 매개변수는 약 1조 개, Gemini 울트라는 1조7500개 수준으로 추정한다. 대규모 언어 모델은 범용성 측면에서는 매우 좋지만 클라우드 서비스 및 전력 비용에 대한 부담이 크다. 역설적으로 LLM의 한계는 그 파라미터 규모와 계산 요구량에 있다. 반면, 소규모 언어모델은 파라미터 매개변수가 상대적으로 작은 언어 모델로 대규모 언어 모델에 비해 더 작은 파라미터 크기와 계산 비용을 가지고 있어 기업의 특정 업무 최적화 및 스마트폰, 노트북과 같은 모바일 기기에서 작동가능하다.

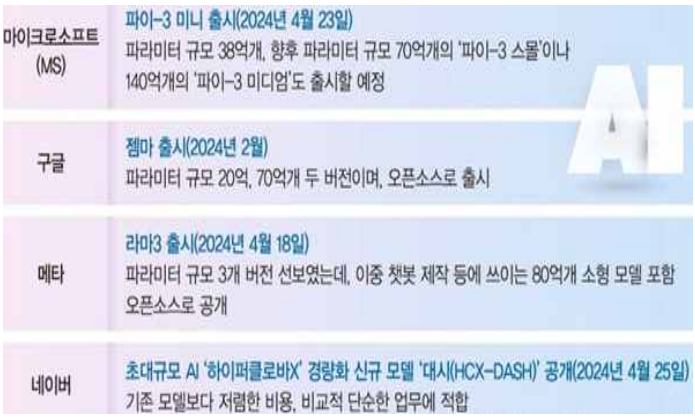

그림 2. 최근 경량화 AI모델 현황

최근 마이크로소프트(MS)는 경량 AI 모델 파이-3를 공개했다. 파이-3 미니의 파라미터는 3.8B 개로, 향후 출시될 '파이-3 스몰(7B)'이나 '파이-3 미디엄(14B)'보다도 매개변수 규모가 작다. 구글은 지난 2월 간단한 챗봇이나 언어 관련 작업에 유용한 파라미터 젬마2B와 7B를 출시했다. 메타는 라마3를 출시하며 파라미터 70B 모델과 함께 챗

봇과 코딩을 지원한다. 빅테크뿐만 아니라 스타트업들도 굉장히 두각을 나타내고 있다. 업스테이지는 자체 sLLM '솔라 미니'를 아마존웹서비스(AWS)를 통해 출시했다. 또한, 유럽의 오픈AI 불리는 미스트랄도 지난 10월에 7.3B 개 매개변수를 갖춘 기업용 소형 AI 모델을 출시했다.

3. sLLM 경량화 연구 동향

연산에 필요한 컴퓨팅 자원은 모두 비용이다. 모델의 정밀도를 낮추어 속도와 메모리 사용량을 절감하는 양자화(Quantization) 연구가 중요하다. 생성 언어모델의 연산 과정에서 적용할 수 있는 8비트 양자화 연구[3]는 sLLM의 확산에 큰 영향을 준다. 양자화는 모델의 가중치와 활성화 함수를 낮은 비트 정밀도로 변환하여 모델의 크기를 줄이고, 연산 속도를 빠르게 하는 기술이다. 예를 들어, 32비트 부동 소수점 형태의 매개변수를 8비트 정수로 변환함으로써 모델의 메모리 사용량을 줄이고, 추론 속도를 향상시킬 수 있다. 이는 특히 자원이 제한된 환경에서 모델을 배포할 때 유용하다[4]. 또한, 미세조정 학습 과정에서는 모델의 일부분만 갱신하여 높은 성능을 내고, 적은 메모리로도 학습 가능한 방법으로 LoRA(Low-Rank Adaptation)[5]와 이를 양자화한 QLoRA 기법이 있다.

3. sLLM 기술의 주의사항

sLLM은 아직 초기 단계의 기술이며, sLLM 기술을 책임감있게 개발하고 활용하기 위해서는 해결할 여러사항이 있다. 편향을 줄이기 위해 다양성이 풍부한 데이터셋을 사용하거나, 설명 가능성을 높이기 위해 모델의 의사 결정 과정을 더 투명하게 만들 수 있다. 또한, 데이터 프라이버시를 보호하기 위해 데이터 암호화, 접근 제어, 데이터 최소화 원칙 등을 적용할 수 있다.

3.1. 모델 편향

sLLM은 학습 데이터의 편향을 반영할 수 있다. 이는 모델이 특정 그룹이나 개인에 대해 부정적이거나 긍정적인 편향을 가질 수 있음을 의미한다. 예를 들어, 데이터셋에 특정 성별이나 인종에 대한 부정적인 예시가 많다면, 모델은 그러한 편향을 학습하여 잘못된 결과를 내놓을 수 있다. 이러한 편향은 모델의 성능 저하뿐만 아니라 사회적 차별 문제를 야기할 수 있음에 주의한다.

3.2. 설명 가능성 부족

sLLM은 복잡한 내부 구조를 가지고 있어, 모델이 어떻게 특정 결정을 내렸는지 설명하기 어렵다. 이는 사용자가 모델의 의사 결정 과정을 이해하고 신뢰하는 데 어려움을 겪게 할 수 있으며, 잘못된 결정이나 오류가 발생했을 때 그 원인을 파악하기 어렵게 만든다.

3.3. 데이터 프라이버시

sLLM은 개인 정보를 포함할 수 있는 데이터를 학습하는 과정에서 프라이버시 문제가 발생할 수 있다. 모델이 개인의 신원, 건강 상태, 금융 정보 등 민감한 데이터를 학습하게 되면, 이 정보가 유출되거나 부적절하게 사용될 위험이 있다. 개인 정보 보호를 위한 철저한 조치가 필요하다.

4. sLLM 활용 분야

sLLM은 스마트폰, 태블릿 등 모바일 기기에서 사용되는 AI 서비스에 활용된다. 음성 명령 인식, 개인화된 추천, 번역 등 다양한 모바일 AI 기능에 활용될 수 있다. sLLM은 다양한 IoT 기기의 데이터를 처리하고 분석하여 사용자에게 편리한 서비스를 제공할 수 있다. 음성 제어, 환경 감지, 자동화 등 다양한 IoT 기능에 활용가능하다. 또한, sLLM은 클라우드로 데이터를 전송하지 않고도 실시간으로 데이터를 처리하고 분석할 수 있어 지연 시간을 줄이고 개인 정보보호를 강화할 수 있다. sLLM은 사용자의 검색 기록, 구매 이력, SNS 활동 등 다양한 데이터를 기반으로 사용자에게 맞춤형 콘텐츠, 상품, 서비스를 추천할 수 있어서 쇼핑, 음악, 영화 등 다양한 서비스 분야에서 활용될 수 있다.

5. 결론

sLLM은 LLM의 핵심 성능을 유지하면서 경량화하여 실제 서비스에 적용하기 쉬운 장점을 가지고 있다. 향후 sLLM은 모바일 AI, IoT, 엣지 컴퓨팅, 개인화 서비스 등 다양한 분야에서 활용될 것으로 예상된다. sLLM은 인공지능을 더욱 다양한 분야에 적용하고 사람들의 삶을 더욱 풍요롭게 만들 수 있는 잠재력을 가지고 있다. sLLM은 아직 초기 단계의 기술이며, sLLM 기술을 책임감 있게 개발하고 활용하기 위해서는 윤리적, 사회적 문제를 해결해야 한다. 향후 지속적인 연구개발을 통해 sLLM의 성능과 효율성이 더욱 향상되고, 다양한 분야에 적용되는 범위가 확대될 것으로 기대된다.

참 고 문 헌

[1] Hailin Chen, et al, "ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?", 2024, arXiv:2311.16989

[2] https://brunch.co.kr/@b2439ea8fc654b8/18

[3] T. Dettmers et al., "LLM.int8(): 8-bit matrix multiplication for transformers at scale," arXiv preprint, CoRR, 2022, arXiv: 2208.07339.

[4] https://blog-ko.superb-ai.com/lightweighting-llm-with-peft/

[5] H. Liu et al., "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," arXiv preprint, CoRR, 2022, arXiv: 2205.05638.

[6] T. Dettmers et al., "QLoRA: Efficient finetuning of quantized LLMs," arXiv preprint, CoRR, 2023, arXiv: 2305.14314.