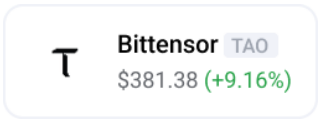
korbit

MESSARI

소형 AI 모델의 놀라운 가능성

Doing More with Less: The Surprising Case for Smaller AI Models

관련 자산



Prologue

AI 개발이 가속화되면서 시장의 관심은 대형 모델과 이를 뒷받침하는 인프라에 집중되어 왔다. 그러나 이번 메사리 리포트에서 강조한 바와 같이 AI의 활용에서 중요한 것은 모델의 크기만이 아니라 그 적용 방식과 통합 과정이다. 실제 애플리케이션에서 소형 AI 모델은 더 다양한 성과와 효율성을 보여주고 있다. 이 모델들은 비용을 절감하고 유연한 개발을 가능하게 하며, 인센티브화(incentivization) 측면에서 크립토와의 시너지도 발생할 수 있다. 대형 모델에 의존하기보다 소형 모델을 시스템에 통합하여 정교한 작업을 처리하는 것은 AI의 잠재력을 극대화할 수 있을 뿐만 아니라 새로운 패러다임으로의 전환을 예고한다. 이번 메사리 리포트는 이러한 AI 개발의 새로운 방향성을 제시하며 소형 AI 모델이 가져올 혁신 가능성에 주목하고 있다.

2024년 10월 18일

코빗 리서치센터장 **최 윤 영**

코빗 리서치센터장 **김 민 승**

핵심 내용

- 지금까지 모든 관심은 주로 AI 스택의 하위 단계에 집중되었으며, 여기에는 OpenAI와 Anthropic 같은 저명한 AI 연구소와 Nvidia 같은 하드웨어 제조업체들이 포함되어 있다.
- 이러한 AI 스택의 하위 단계에 대한 관심과 자본의 집중으로 인해 애플리케이션 레이어에 축적되고 있는 잠재력이 가려지고 있다.
- 앞으로 몇 달 동안 이 애플리케이션 레이어에서 실험이 계속 증가함에 따라, 애플리케이션에 AI 모델을 통합하는 개발자들은 (매우 구체적이거나 특수한 기능을 갖춘) 소형 모델을 사용하는 것이 더 관리하기 쉽고 유연한 AI 시스템을 구축하는 데 도움이 된다는 것을 알게 될 것이다.
- 소형 AI 모델의 사용은 분산형 트레이닝, 로컬 추론, 데이터 세트 수집 및 생성 등 AI x 크립토 스택 내의 여러 영역에서 긍정적인 신호를 보인다.

2022년 말로 돌아가 보자. 전 세계는 OpenAI의 챗GPT에 내장된 마법 같은 속성을 처음 경험하고 있었다. 대부분의 초기 실험은 새롭지만 혁신적인 기술에서 흔히 볼 수 있는 전형적인 패턴을 따랐는데, 주로 재미있는 장난감으로 사용되고 이해되었다.

현재로 빠르게 넘어오면 챗GPT가 일으킨 불꽃은 최초의 인공 일반 지능(AGI, artificial general intelligence)을 개발하기 위한 노력을 지원하는 거대한 자금 확보 경쟁으로 이어졌다. 이 목표를 염두에 두고 모든 관심은 차세대 조 단위 파라미터 모델을 개발하는 대규모 AI 연구소(OpenAI와 Anthropic 등)와 하드웨어 제조업체(Nvidia 등)에 집중되었다.

AI 연구소와 하드웨어 회사는 AI 스택의 하위 단계를 대표한다. 이러한 레이어들이 결합되어 AI 에이전트, 애플리케이션, 시스템이 등장하는 빌딩 블록을 형성한다. 이러한 하위 단계 스택에 관심과 자본이 집중되면서 애플리케이션 레이어에 잠재된 가능성이 가려졌다. 비교적 단순한 AI 에이전트인 챗GPT가 보여준 것처럼 이러한 모델에서 비롯되는 진정한 마법은 다른 소프트웨어 시스템과 통합되어 응집력 있는 제품을 만들 때 느낄 수 있다.

보다 일반적으로 순수 AI 모델을 툴, 오케스트레이션 소프트웨어(orchestration software), 비즈니스 로직, 그리고 추가적인 AI 모델들과 결합하면 해당 애플리케이션은 버클리 AI 연구 그룹이 명명한 것처럼 AI 시스템 또는 복합(compound) AI 시스템으로 간주할 수 있다. 그들이 언급했듯이, 이러한 시스템은 단일 AI 모델만으로는 얻을 수 없는 놀라운 결과를 달성할 수 있다.

더 많은 개발자들이 애플리케이션에 AI 모델을 통합하는 실험을 하면서, (매우 구체적이거나 특수한 기능을 갖춘) 소형 모델일수록 더 관리하기 쉽고 유연한 AI 시스템을 만들 수 있을 것이다. 비용 절감만으로도 소형 모델 사용을 탐구할 매력적인 이유가 된다. OpenAI의 대형 GPT-4o 모델을 사용하는 것은 GPT-4o 미니 모델보다 약 30배 더 비싸다.

소형 모델 사용이 계속 증가하는 세상에서는 탈중앙화(decentralized) 모델 트레이닝, 로컬 추론, 데이터 수집 인센티브화(incentivization)와 같은 분야에서 긍정적인 2차 효과가 발생할 가능성이 있으며, 이는 AI x 크립토 스택 내 많은 팀들이 집중하고 있는 분야이다.

AI 시스템과 소형 모델

위에서 설명한 이러한 AI 시스템으로의 전환은 독립형(standalone) AI 모델 사용으로 자연스럽게 발전해 나가는 과정이다. 일반적으로 AI 모델 자체는 최종 사용자가 원하는 제품이 아니다. 오히려 가치를 창출하는 것은 소프트웨어 시스템 전체(즉 AI 애플리케이션)이다.

일반적인 애플리케이션 개발과 마찬가지로 AI 모델 하나 또는 여럿을 비즈니스 로직 및 필요한 도구와 함께 패키징하려면 신중한 설계가 필요하며 많은 테스트 사이클, 반복, 배포 과정이 필요할 수 있다. 따라서 더 작고 특화된 모델이 아키텍처 관점에서 더 유리할 수 있다.

소형 AI 모델의 장점

대형 언어 모델(LLM, large language models)의 스케일링 법칙은 지금까지는 유효했다. 컴퓨팅 예산과 트레이닝 데이터 세트의 크기와 함께 모델의 전체 크기를 늘리면 일반적으로 성능이 더 뛰어난, 즉 "더 지능적인" 모델이 만들어졌다. 그러나 이러한 대형 모델의 성능 향상에는 작은 모델에 비해 몇 가지 트레이드오프가 따른다.

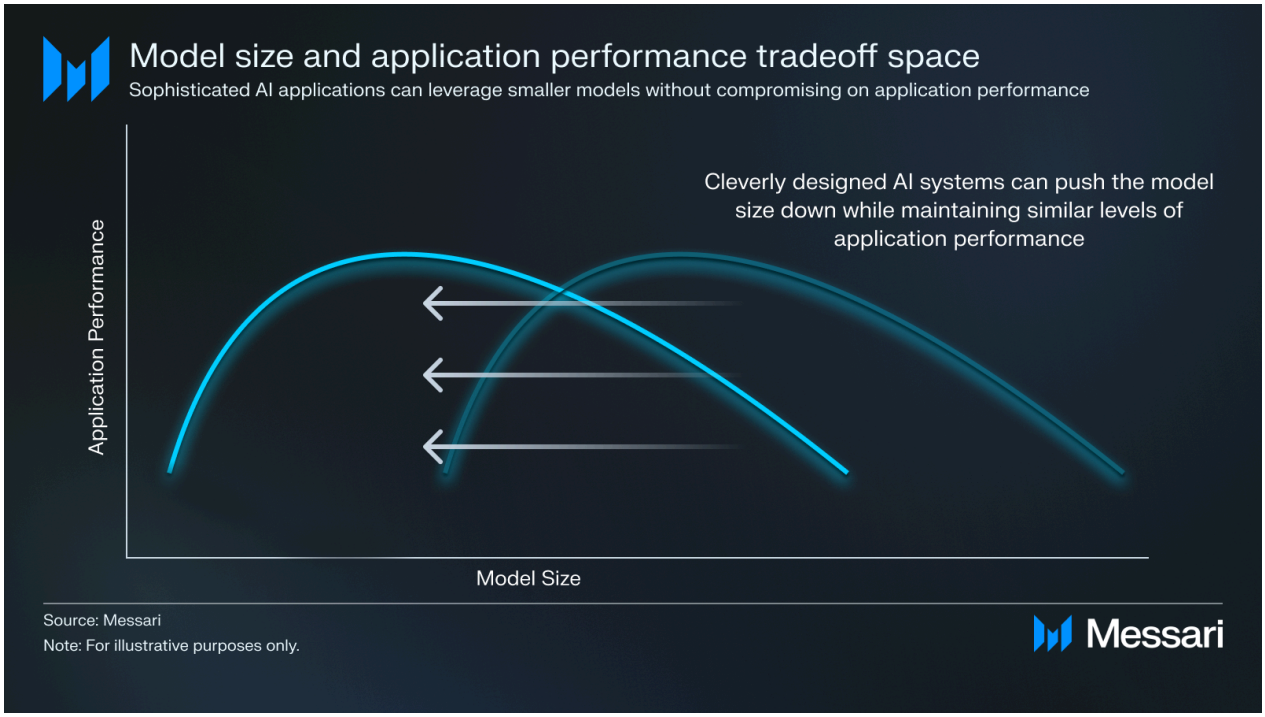
모델 트레이닝 및 추론의 트레이드오프

소형 모델일수록 최첨단 LLM보다 더 적은 컴퓨팅 예산으로 더 빠르게 학습할 수 있다. 예를 들어, 스케일링 스펙트럼의 끝자락에 있는 Meta의 Llama 3.1 모델 세트는 방대한 트레이닝 데이터 세트(15조 개 이상의 토큰)를 정교한 Nvidia H100 GPU 클러스터와 함께 활용했다. 메타의 16,000개 H100 클러스터를 사용한 가장 큰 모델을 학습시키는 데는 몇 달은 아니더라도 몇 주가 걸렸을 것이다. 반면에 더 작은 모델은 더 적은 수의 GPU로 더 짧은 시간에 더 적은 양의 데이터로도 학습이 가능하다. 빠르게 출시하고 다양한 제품 디자인을 반복하는 데 중점을 두는 애플리케이션 개발자에게는 앱 내에서 좁은 범위에 사용되며 광범위한 일반 지능을 요구하지 않는 소형 모델이 매력적인 선택이 될 수 있다.

모델 추론(inferencing)이란, 학습되고 애플리케이션에 통합된 모델에 효과적으로 질의하는(querying) 것을 의미한다. 모델의 응답 시간, 즉 지연 시간(latency)은 AI 애플리케이션이 실제 사용자 요청을 처리하는 프로덕션 단계에서 중요한 지표가 된다. 일반적으로 지연 시간을 최소화하면 전반적인 사용자 경험(UX)이

개선된다. 예를 들어, 챗GPT가 각 쿼리에 응답하는 데 몇 분, 또는 수십 초가 걸린다면 UX가 얼마나 열악하게 느껴질지 상상해 보라. 더 작은 모델은 전반적인 연산 자원이 덜 필요하기 때문에 더 큰 모델보다 추론 요청을 더 빠르게 처리할 수 있다.

이러한 설계 결정은 애플리케이션 개발자가 해결해야 할 트레이드오프의 범위를 형성한다. 애플리케이션의 요구 사항을 충족하는 (지연 시간 및 비용 측면에서의) 성능을 제공하면서, 모델을 얼마나 작게 만들 수 있을까?



지금까지 성능을 크게 향상시키고자 하는 개발자가 선택할 수 있는 최적의 방법은 대형 최첨단 AI 모델의 다음 버전을 기다리는 것이었다. 이러한 대형 모델은 더 많은 기능을 제공하지만 그만큼 컴퓨팅 리소스와 지연 시간 비용이 증가했다. 많은 사람들에게 이러한 대형 모델에 의존하는 것이 원하는 애플리케이션 UX를 만드는 유일한 실질적 경로처럼 보였다. 그러나 모든 애플리케이션이 이러한 대형 모델의 모든 기능을 필요로 하는 것은 아니기 때문에 효율성 측면에서 이러한 접근 방식은 사실상 과잉이며, 이로 인해 모델의 기능과 애플리케이션의 실제 요구 사항 사이에 불일치가 발생했다.

현재 변화가 일어나고 있다. 더 작은 AI 모델은 실제 프로덕션 환경에 배포할 수 있을 만큼 충분히 강력해지고 있다. 복잡한 비즈니스 로직과 툴 사용, 함수 호출, 검색 증강 생성(RAG, retrieval-augmented generation) 시스템, 파인 튜닝(fine-tuning) 및 기타 소형 AI 모델과 같은 기술을 결합하면 이러한 AI 시스템은 대형 모델을 활용하는 것에 필적하거나 그 이상의 결과를 만들어낼 수 있다.

AI x 크립토에 미치는 영향

AI 시스템과 애플리케이션에 더 작은 AI 모델을 도입하면 탈중앙화된 트레이닝, 로컬 추론, 인센티브화된(incentivized) 데이터 수집 등 AI x 크립토 스택 내의 여러 분야에 긍정적인 영향을 미칠 수 있다.

탈중앙화 및 분산형 트레이닝

최근 탈중앙화 및 분산형 트레이닝의 획기적인 발전으로 이 개념이 AI의 주류로 부상했다. Prime Intellect와 Nous Research는 서로 다른 기술을 사용하여 지리적으로 분산된 컴퓨팅 클러스터를 이용해 AI 모델을 훈련하는 것이 가능하다는 점을 입증했다. 이러한 연구 결과가 발표되기 전까지는, 탈중앙화된 트레이닝은 현실적으로나 경제적으로 달성 불가능한 것으로 여겨졌다.

인상적인 초기 결과에도 불구하고, 이러한 트레이닝 방법을 확장하여 순수 모델 크기 측면에서 OpenAI 및 Anthropic과 같은 AI 연구소에서 생성하는 모델(예: 1조 개 파라미터 모델)과 동등한 수준의 모델을 만들기 위해서는 더 많은 연구와 엔지니어링 작업이 필요하다.

그러나 (Prime Intellect와 Nous가 수십억 개의 파라미터를 가진 더 작은 모델을 실험한 것처럼) 소형 모델을 활용하는 경우, 이러한 새로운 분산형 트레이닝 방법을 시스템에 통합하면 Gensyn 및 Prime Intellect와 같은 탈중앙화 트레이닝 프로토콜을 활용할 수 있다.

로컬 추론

텍스트, 이미지, 동영상 등 최신 생성형 AI 모델의 대부분의 사용자는 호스팅 서비스를 통해 해당 모델과 상호작용한다. 예를 들어, 기존의 전통적인 클라우드 아키텍처와 마찬가지로 OpenAI는 챗GPT 애플리케이션을 효과적으로 호스팅 및 실행하며, 개발자가 자신의 모델 세트와 통합할 수 있도록 API 엔드포인트를 제공한다.

호스팅 서비스가 사용자에게 제공하는 편의성은 아무리 강조해도 지나치지 않다. 하지만 다음과 같은 단점도 있다:

- 블랙박스 모델 – 오늘 사용하는 모델이 내일은 다를 수 있다. 사용자 입장에서 모델은 서비스 제공업체가 사용자 모르게 변경할 수 있는 블랙박스와 같다. 이로 인해 예상치 못한 동작이 발생하거나, 더 높은 품질의 모델에 대해 비용을 지불했음에도 불구하고 더 낮은 품질의 모델로 대체되는 상황이 발생할 수 있다.
- 개인정보 보호 – 서비스를 운영하는 주체는 모델을 통해 전달되는 모든 데이터를 볼 수 있다. 이로 인해 사용자가 자신의 쿼리를 비공개로 유지할 수 있는 권한이 사라진다.

엑소 랩스(Exo Labs) 팀은 사용자가 로컬 디바이스(예: 노트북 또는 스마트폰)에서 오픈소스 모델을 실행할 수 있는 사용하기 쉬운 SDK를 개발하여 이러한 문제를 해결하고 있다. 일반적으로 휴대폰과 같은 로컬 엣지 디바이스(local edge device)에는 AI 모델과 같이 연산량이 많은 소프트웨어를 실행하는 데 필요한 하드웨어가 부족하다. 엑소 랩스 SDK는 여러 장치를 연결해 더 뛰어난 성능의 단일 하드웨어처럼 작동하도록 지원한다. 크립토 관점에서 보면, 로컬 모델을 온체인 스마트 콘트랙트 작업을 실행(trigger)하도록 구성할 수 있다.

여전히 가장 성능이 뛰어난 GPU에는 미치지 못하지만, 이러한 소프트웨어를 통해 사용자는 자신의 기기에서 소형 오픈 소스 모델을 실행할 수 있다. 간단한 AI 애플리케이션(예: 챗GPT와 같은 챗봇)의 경우, 로컬에서 실행하면 위에서 설명한 블랙박스 및 개인정보 보호 문제를 모두 해결할 수 있다.

트레이닝 데이터 인센티브화와 혁신

데이터는 AI 모델의 성격이나 일반적인 행동을 형성하는 데 핵심적인 요소이다.

예를 들어, character.ai와 유사한 AI 컴패니언 애플리케이션을 만들고 있는 Dippy를 생각해 보자. Dippy Bittensor 서브넷은 롤플레이 LLM의 생성을 인센티브화(incentivize)하며, 이렇게 생성된 LLM은 애플리케이션에 추가로 통합된다. EQ 벤치마크는 롤플레이 LLM의 품질을 결정하는 주요 지표 중 하나이다. 이는 모델의 감성 지능을 효과적으로 측정하기 위한 것으로, 이는 AI 컴패니언에게 매우 중요한 요소이다. Dippy 팀은 EQ 벤치마크 측정에 사용되는 데이터 세트를 혁신함으로써 메타 및 OpenAI에서 생성한 비슷한 크기의 모델(약 70억~80억 개의 파라미터)을 능가하는 성능을 보였다.

보다 광범위하게는 Bittensor와 같은 크립토 시스템을 활용하여 특정 데이터 세트를 수집하고 생성하도록 인센티브화(incentivize)하는 것이 새로운 패턴으로 떠오르고 있다. Dippy의 서브넷 외에도, Macrocosmos의 서브넷 13은 다양한 소스에서 데이터를 스크랩하는 데 중점을 두고 있다. 앞으로 이 서브넷은 온디맨드 방식으로 데이터를 스크랩하거나, 데이터를 소싱하기 위해 누구나 서브넷의 리소스를 지시할 수 있게 하는 자연스러운 확장이 이루어질 것이다. 이는 매우 특정한 데이터 세트에 의존하는 더 작고 특화된 모델을 활용하는 사람들에게 매력적인 옵션이 될 것이며, 크립토 인센티브 메커니즘이 이 서브넷의 운영에 중요한 역할을 할 것이다.

글을 맺으며

의심할 여지 없이, 차세대 프론티어 모델(예: GPT-5)과 고성능 컴퓨팅 칩의 출시는 계속해서 대중의 관심을 끌 것임에 틀림없다(그럴 만한 이유도 충분하다). 그러나 많은 사람들이 챗GPT에서 경험했듯이, 이러한 모델의 진정한 영향력은 그것들이 애플리케이션이나 제품에 영리하게 통합될 때 비로소 느껴진다.

개발자들이 단순한 챗봇 스타일의 AI 애플리케이션을 넘어서기 위해 노력함에 따라, 이러한 시스템에 내장된 소형 모델은 매력적인 솔루션이 될 수 있다. 일반적으로 소형 모델은 더 큰 모델보다 비용 효율적일 뿐만 아니라 모듈식 소프트웨어로 작동할 수 있어 전체 제품을 더욱 유연하게 만들 수 있다. 탈중앙화된 트레이닝 및 컴퓨팅 프로토콜, 로컬 추론 프로젝트 개발자, 데이터 세트 수집 프로토콜 모두 이러한 소형 모델의 사용 증가에 따라 이점을 누릴 수 있다.

어쩌면 미래의 어느 시점에 AI 연구소가 인공 일반 지능(AGI)을 만드는 데 성공하여 “소형” 모델을 사용할 필요가 없어질지도 모른다. 그때까지는 AI 애플리케이션 개발자들은 더 적은 자원으로 더 많은 일을 할 수 있는 방법을 계속 찾아낼 것이다.

By Seth Bloomberg

SEP 6, 2024 · Pro

원문 링크

https://messari.io/report/doing-more-with-less-the-surprising-case-for-smaller-ai-models