DeepSeek 모델 기술 분석

김성표[1], 박삼준[2], 유이주[1]

[1]국방과학연구소, [2]국방기술진흥연구소

spkim35@add.re.kr, samnjoonpark@krit.re.kr, youyiju@add.re.kr

An Analysis of AI Technologies Applied to DeepSeek Models

Kim Seong Pyo[1], Park Samjoon[2], You Yiju[1]

[1]*ADD,* [2]*KRIT*

요 약

최근 Deep Seek는 저비용으로 고성능의 대규모 언어모델(LLM)을 개발하여 주목받고 있고 다양한 버전의 모델을 출시하며 기술혁신을 이어가고 있다. 본 고에서는 Deep Seek가 개발한 저비용 고성능 모델에 적용된 AI 기술(Mixture of Experts, Chain of Thought, Distillation 등)을 분석해 보고 그 기술 적용의 의미를 해석해 보았다.

Ⅰ. 서 론

Deep Seek는 중국의 AI 스타드업으로 저비용으로 고성능의 대규모 언어모델(LLM)을 개발하여 주목받고 있다. 이들은 DeepSeek-LLM (Jan. '24)[1], DeepSeek-V2 (Jun '24)[3], DeepSeek-V3 (Dec '24)[4], DeepSeek-R1 (Jan '25)[5] 등 다양한 버전의 모델을 출시하며 기술혁신을 이어가고 있다. 따라서 Deep Seek가 어떻게 저비용 고성능 LLM 모델을 개발할 수 있었는지 적용된 AI 기술을 분석하고 그 기술 적용의 의미를 해석해 볼 필요가 있다.

Ⅱ. DeepSeek 모델 버전별 개요 및 기술

DeepSeek 모델의 각 버전은 성능 향상과 효율성 증대를 목표로 다양한 기술을 도입하였다. 주요 모델 버전과 그에 적용된 핵심 기술들을 표 1에 요약하였다.

표 1 DeepSeek 모델 버전별 적용된 주요 기술 및 특징

버전	출시 시기	파라미터 수		주요 기술 및 특징
		전체	활성화	
LLM	2024.01	67B	전체 사용	범용 텍스트 생성, 2조 토큰 학습, 오픈소스
MOE	2024.01	16B	2.7B	MoE 구조 도입하여 효율성 향상
V2	2024.05	236B	21B	MLA, DeepSeekMoE, 다양한 변형 모델 출시
V3	2024.12	671B	37B	MLA, DeepSeekMoE, FP8 도입, 보조 손실 없는 로드 밸런싱
R1/ R1-Zero	2025.01	671B	37B	강화 학습 기반 학습, GRPO, Chain of Thought, 수학 및 논리 추론 최적화

Ⅲ. Deep Seek 모델에 적용된 AI 기술

1. MoE(Mixture-of-Experts)

MoE는 다양한 지식 영역에 대해 각각 특화된 매개변수를 가진 개별 영역(전문가)으로 모델을 논리적으로 분할하는 설계 기법이다. MoE 적용시 각 토큰은 가장 관련성이 높은 전문가에게만 라우팅된다. 따라서, MoE는 완전히 밀집된 접근 방식(Fully dense approach)에 비해 필요한 계산을 크게 줄일 수 있다.

이 접근 방식은 각 전문가가 서로 다른 GPU에 상주할 수 있기 때문에 장치 간 통신을 제한(예: Device-limited routing[2])함으로써, MoE는 메모리나 데이터 전송 비용을 크게 들이지 않고 매우 큰 수의 매개변수로 효과적으로 확장(scale)할 수 있다.

Deep Seek는 MoE 구조를 도입하여 전체 모델 파라미터 수는 많지만, 실제 연산 시에는 일부 전문가 네트워크만 활성화되도록 설계했다. 예를 들어, DeepSeek-V2 모델은 총 2,360억 개의 파라미터를 보유하지만, 추론 시에는 약 210억개만을 사용하고[2] DeepSeek-V3 모델은 총 6,710억 개의 파라미터를 보유하지만, 추론 시에는 약 370억개만 사용한다[3]. 이러한 방식은 필요한 전문가만을 선택적으로 활성화하여 연산량과 메모리 사용을 줄여 하드웨어 부담을 낮추는 데 기여한다.

2. Multi-Head Latent Attention (MLA)

MLA는 기존의 Multi-Head Attention보다 효율적인 구조로, Key-Value(KV) 캐시를 잠재 백터(Latent vector)로 압축하여 메모리 사용량을 크게 줄일 수 있다[2].

DeepSeek-LLM과 비교했을 때, DeepSeek-V2는 훨씬 더 강력한 성능을 달성하며, 동시에 훈련 비용의 42.5%를 절감하고 KV 캐시를 93.3% 줄이며 최대 생성 처리량(Throughput)을 5.76배까지 향상시켰다고 한다[2]

3. CoT(Chain-of-Thought)

CoT은 언어 모델이 답을 찾아가는 과정에서 단계별 추론 과정(Step-by-step reasoning)을 따르도록 유도하는 방식을 사용하여 복잡한 작업에서 언어 모델의 성능을 향상시키기 위해 사용되는 기술이다. 이 기술은 수학 문제, 복잡한 이해 문제 또는 주어진 데이터로부터 추론을 요구하는 문제와 같이 여러 단계나 추론의 레이어가 필요한 문제에 특히 유용하다. 예를 들어, 수학 문제의 해답을 간단히 묻는 대신에, 문제를 접근하는 방법의 분해를 프롬프트에 포함시켜 모델이 이러한 사고 과정을 반영하도록 유도할 수 있다.

그림 1에 일반(Standard) Prompting과 CoT Prompting를 비교한 예시를 보여주고 있다[5]. 그림에서 우측의 파란색으로 칠해진 부분처럼 수학 문제를 푸는데 사용한 로직을 차근 차근 단계별로 풀어서 예

시를 제시해 주었더니, LLM 모델이 Output을 생성할 때 역시 차근 차근 단계를 밟아서 문제 풀이 로직을 설명해 가면서 정확하게 답을 생성해 내었다.

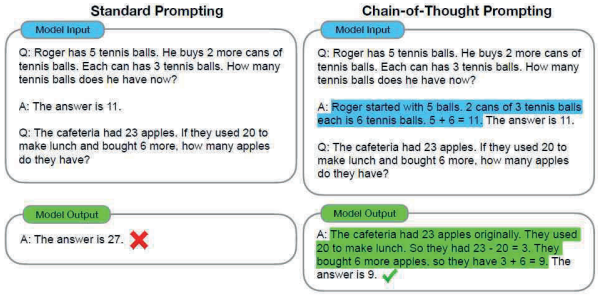

그림 1 Chain-of-thought prompting 사례

CoT를 적용한 DeepSeek 모델이 복잡한 문제를 단계적으로 해결하고, 정확하고 신뢰할 수 있는 결과를 제공하는 데 기여했다고 볼 수 있다.

4. Model Distillation

Model distillation은 크고 복잡한 모델(교사 모델)에서 얻은 지식을 더 작고 단순한 모델(학생 모델)로 옮기는 방법이다. 모델 증류를 통해서 더 큰 모델의 성능을 대부분 유지하면서도 계산 능력, 메모리 사용량, 추론 속도 측면에서 효율성이 향상된 더 컴팩트한 모델을 개발할 수 있다.

Deep Seek는 대형 모델의 성능을 유지하면서도 경량화된 모델을 생성하기 위하여 Distillation 기법을 사용하였다.

IV. 종합 해석

1. 성능 측면

Deep Seek가 어떻게 저비용 고성능 LLM 모델을 개발할 수 있었는지 DeepSeek 모델에 적용된 AI 기술을 분석하였다. 이러한 기술들은 DeepSeek 모델에만 적용된 것은 아니라 기존의 다른 모델에도 적용되고 있다. 예를 들면, DeepSeek-R1의 CoT(Chain of Thought)는 DeepSeek가 이룬 기술적 돌파구라고 평가하고 있다. 하지만 실제로 DeepSeek-R1과 ChatGPT-o1은 모두 CoT를 적용하고 있다. 다만, DeepSeek-R1은 CoT 과정을 UI로 보여주는데 반하여 ChatGPT-o1은 내부적으로 처리하고 UI로 보여주지 않는다는 차이가 있을 뿐이다. 즉, UI 문제일 뿐이지 기술적 혁신이라고 보기는 어렵다고 볼 수 있다..

그러나 Deep Seek는 이러한 기술을 기존과 다른 최적화 기법을 적용하여 효율을 극대화하고 성능을 향상시킨 것으로 보인다. 그림 2에서 보는 바와 같이 DeepSeek 모델은 다양한 벤치마크에서 우수한 성능을 보였다. 특히 코드 생성, 복잡한 수학문제 해결, 단계적 추론에서 OpenAI-o1과 유사하거나 뛰어난 성능을 기록하였다.

이는 GPT-Style Dense Model과 같은 단일한 기술흐름에 의존하던 LLM 시장에 새로운 기술적 대안을 제시한 것으로 보인다.

그러나, 성능을 최적화한 것과 효율성을 최적화한 것은 큰 차이가 있다. DeepSeek의 추론 모델 R1이 OpenAI의 추론 모델인 o1의 성능에 견줄만 하다고 하지만 OPenAI는 성능이 더 우수한 o3를 이미 시연했다. DeepSeek 모델은 효율성에서 앞섰다고 할 수 있으나 전반적인 성능에서 앞선 것은 아니라고 볼수 있다.

2. 비용 측면

Deep Seek가 주장하는 저비용으로 대규모 언어 모델(LLM)을 개발할 수 있었던 배경에는 표 2에 제시한 바와 같이 효율성을 극대화한 여러 기술적 전략이 있었던 것으로 보인다.

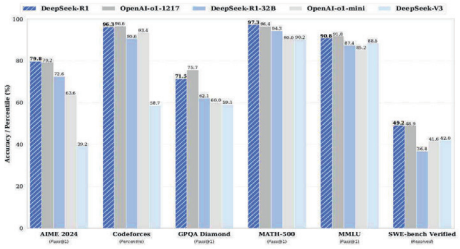

그림 2 Perfromance of DeepSeek-R1

그러나, DeepSeek 모델도 기본적으로 Transformer 아키텍처를 사용하고 있기 때문에 복잡한 행렬계산이 필요하고 고속으로 데이터를 메모리에 저장해야 하는 특성이 있어 저비용 LLM이라고 할 수 있을지는 추가 검증이 필요해 보인다. 특히, $5,58M에 DeepSeek-V3를 훈련시켰다고 하나 그것은 마지막 훈련 과정(Final training run)에 소요된 비용이며 인프라스트럭쳐 같은 핵심 비용은 제외한 금액이다. 그리고 OPenAI도 o3-mini를 무료로 서비스를 제공하고 있다.

표 2 효율성을 극대화한 기술적 전략

전 략	하드웨어 관점의 이점
MOE 아키텍처	불필요한 연산감소로 자원 절약
MLA	메모리 사용량 감소로 추론 속도 향상
FP8 연산	메모리/대역폭 절약, 저사양 GPU 사용 가능
저사양 GPU 사용	고급 하드웨어 없이도 대규모 모델 훈련 가능
Distillation	경량화된 모델로 훈련 비용 및 자원 절약

3. 안전성 측면

Deep Seek의 서버가 중국에 있다는 점에서 발생할 수 있는 문제점이 있다. 그것은 중국정부의 정보통제 및 사이버보안 법률에 따른 데이터 프라이버시, 콘텐츠 검열 및 제한, 서비스의 투명성 부족, 지속적 접근성 문제 등이 있을 수 있다.

V. 결론

Deep Seek의 논문을 분석해 보면 MOE 아키텍처, MLA, FP8 연산, CoT, Distillation 등 기술을 기존과 다른 최적화 기법을 적용하여 비용 및 메모리 효율성을 높이고 추론 성능을 향상시킨 것으로 보인다. 그러나 일정부분 저비용 효과를 거둘 수 있을 것으로 보이나 DeepSeek 모델도 기본적으로 Transformer 아키텍처를 사용하고 있기 때문에 복잡한 행렬계산이 필요하고 고속으로 데이터를 메모리에 저장해야 하는 특성이 있어 기존 타 모델에 비해 수십배 저비용 LLM이라고 할 수 있을지는 추가 검증이 필요해 보인다. 또한, 중국정부의 정보통제 및 사이버보안 법률로 인하여 데이터 프라이버시 문제 등으로 우려를 낳고 있다.

참 고 문 헌

[1] DeepSeek-LLM (Jan '24): Scaling Open-Source Language Models with Longtermism.
[2] DeepSeek-V2 (Jun '24): A Strong, Economical, and Efficient Mixture-of-Experts Language Model.
[3] DeepSeek-V3 (Dec '24): DeepSeek-V3 Technical Report..
[4] DeepSeek-R1 (Jan '25): Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
[5] Jason Wei, et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", NeurIPS 2022, 28 Jan 2022.