GGUF

Fast campus

GGUF와 GGML은 추론을 위한 모델을 저장하는 데 사용되는 파일 형식으로, 특히 GPT(Generative Pre-trained Transformer)와 같은 언어 모델의 맥락에서 사용됩니다. 각각의 주요 차이점, 장단점을 살펴보겠습니다.

GGML(GPT 생성 모델 언어): 조지 게르가노프가 개발한 GGML은 머신 러닝을 위해 설계된 텐서 라이브러리로, Apple Silicon을 비롯한 다양한 하드웨어에서 대규모 모델과 고성능을 구현할 수 있도록 지원합니다.

장점

- 초기 혁신: GGML은 GPT 모델을 위한 파일 형식을 만들려는 초기 시도였습니다.
- 단일 파일 공유: 단일 파일로 모델을 공유할 수 있어 편의성이 향상되었습니다.
- CPU 호환성: GGML 모델은 CPU에서 실행할 수 있어 접근성이 더욱 넓어졌습니다.

단점

- 제한된 유연성: GGML은 모델에 대한 추가 정보를 추가하는 데 어려움을 겪었습니다.
- 호환성 문제: 새로운 기능 도입으로 인해 이전 모델과의 호환성 문제가 종종 발생했습니다.
- 수동 조정 필요: 사용자가 자주 rope-freq-base, rope-freq-scale, gqa, rms-norm-eps와 같은 설정을 수정해야 했는데, 이는 복잡할 수 있습니다.

2023년 8월 21일, GGML(GPT 생성 모델 언어)의 후속으로 도입된 GGUF(GPT 생성 통합 형식)가 출시되었습니다. 이 형식은 언어 모델 파일 형식 분야에서 중요한

진전을 이루었으며, GPT와 같은 대용량 언어 모델의 향상된 저장 및 처리를 용이하게 해줍니다.

GGML의 창시자 게오르기 게르가노프를 비롯한 AI 커뮤니티의 기여자들이 개발한 GGUF는 독립적인 노력으로 보이지만 대규모 AI 모델의 요구사항에 부합하는 것입니다. Facebook(Meta)의 LLaMA(대규모 언어 모델 메타 AI) 모델과 관련된 컨텍스트에서 사용된다는 점은 AI 환경에서의 중요성을 강조합니다. GGUF에 대한 자세한 내용은 여기에서 GitHub 이슈를 참조하고 여기에서 Georgi Gerganov의 llama.cpp 프로젝트를 살펴볼 수 있습니다.

Pros: 장점

- GGML의 한계를 해결합니다: GGUF는 GGML의 단점을 극복하고 사용자 경험을 향상시키기 위해 설계되었습니다.
- 확장성: 이전 모델과의 호환성을 유지하면서 새로운 기능을 추가할 수 있습니다.
- 안정성: GGUF는 급작스러운 변경을 없애고 최신 버전으로의 전환을 용이하게 하는 데 중점을 둡니다.
- 다목적성: 라마 모델의 범위를 넘어 다양한 모델을 지원합니다.

Cons: 단점

- 전환 시간: 기존 모델을 GGUF로 전환하려면 상당한 시간이 필요할 수 있습니다.
- 적응이 필요합니다: 사용자와 개발자는 이 새로운 형식에 익숙해져야 합니다.

요약

GGUF는 유연성, 확장성, 호환성을 강화한 GGML의 업그레이드 버전입니다. 사용자 경험을 간소화하고 llama.cpp를 넘어 더 광범위한 모델을 지원하는 것을 목표로 합니다. GGML은 초기에는 가치 있는 노력이었지만, GGUF는 그 한계를 해결하

여 언어 모델용 파일 형식 개발의 진전을 의미합니다. 이러한 전환은 모델 공유와 사용 효율성을 향상시킴으로써 AI 커뮤니티에 도움이 될 것으로 기대됩니다.