서버 가상화와 컨테이너 기술을 활용한 Ollama 기반 로컬 LLM 환경 구축 및 운영 방안

이용구
한국수력원자력 중앙연구원
e-mail:lee.yongku@khnp.co.kr

A plan to build a local LLM environment based on Ollama using server virtualization and container technology

Yong-Ku Lee
Central Research Institute, Korea Hydro & Nuclear Power Co., Ltd.

요 약

본 연구는 외부 클라우드 서비스 의존도를 줄이고 데이터 보안을 강화하기 위한 로컬 LLM 환경의 필요성에 주목하여, Ollama 기반의 로컬 LLM 운영 아키텍처를 제안한다. 다수의 VM에 Ollama를 설치하고 GPU 자원을 효율적으로 활용하는 분산 시스템을 구축하여, 네트워크 자율성을 확보하면서도 대규모 언어 모델(LLM)을 안정적으로 운영할 수 있는 방법을 제시하였다. 또한, 웹 기반 인터페이스를 통해 사용자가 쉽게 접근할 수 있도록 컨테이너화된 환경을 구성하였으며, 향후 시스템 확장과 최적화를 통해 더 많은 사용자 요구를 수용할 수 있을 것으로 기대된다.

1. 서론

최근 인공지능(AI) 기술의 급속한 발전과 함께, 대규모 언어 모델(LLM)의 중요성이 크게 부각되고 있다. 특히 transformers 계열의 모델은 다양한 비즈니스 문제 해결에 적용될 수 있으며, 대화형 AI, 문서 생성, 번역 등 폭넓은 활용이 가능하다. 이와 함께 외부 클라우드 서비스 의존도를 줄이고, 데이터 보안과 네트워크 자율성을 강화하기 위한 로컬 환경에서의 LLM 활용 요구가 증가하고 있다. 본 연구는 Ollama를 이용한 로컬 환경에서의 LLM 운영 아키텍처와 구축 절차를 제시하고, 성능 평가를 통해 그 효과성을 검증하고자 한다.

2. 본론

2.1 로컬 LLM 환경 구축의 필요성

기존 클라우드 기반 LLM 서비스는 높은 성능을 제공하지만, 네트워크 종속성, 데이터 보안 우려, 비용 등의 문제로 인해 로컬 환경에서의 대안이 요구된다. 특히, 망 분리 환경이나 데이터 보안이 중요한 산업군에서는 클라우드 접근이 제한적이므로, 자체적인 LLM 운영이 필요하다. Ollama는 로컬 환경에서 이러한 문제를 해결할 수 있는 가벼운 LLM 적재 및 운영 솔루션을 제공한다.

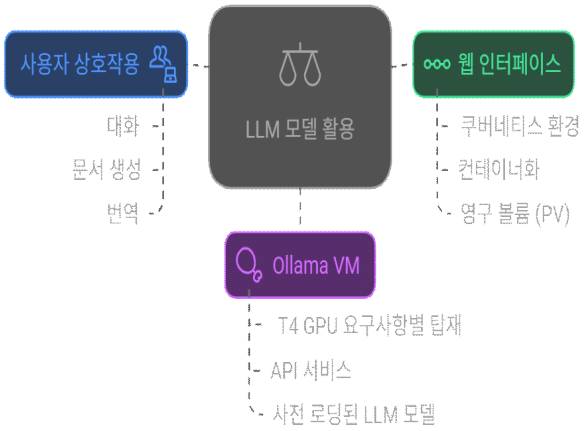
2.2 다수의 Ollama 탑재 VM 구성

본 연구에서 제안하는 시스템 아키텍처는 다수의 VM에 Ollama를 설치하고, 이를 통해 분산된 LLM 환경을 구축하는 방식으로 설계되었다. 각 VM은 정해진 LLM 모델을 사전 로드하여 요청을 처리하도록 구성하며, LLM 모델별 요구사항에 맞도록 GPU(nvidia T4)를 1개에서 4개까지 탑재하여 서비스를 제공한다.

2.3 웹 인터페이스

LLM 모델 사용을 위한 사용자 인터페이스는 쿠버네티스 환경에서 컨테이너 이미지로 배포되며, 사용자에게 LLM 활용을 위한 웹 환경을 제공한다. 이러한 컨테이너 환경은 내부 PV영역에 사용자 정보 및 대화 기록 등을 저장하도록 구성되었다. 또한, 다수 Ollama 서버를 API로 호출하기 위한 연결 설정을 지원한다. 이러한 구성을 통해 향후 사용량 증가 시 쉽게 부하분산 구조를 갖출 수 있다.

2.4 구성도



[그림 1] 시스템 구성도

3. 결론

본 연구에서는 다수의 VM에서 Ollama를 활용한 로컬 LLM 환경 구축 방안을 제안하였다. 제안된 방법은 기업 내 네트워크 자율성과 데이터 보안을 강화하는 동시에, 클라우드 서비스에 의존하지 않고도 요구사항에 맞는 LLM을 효과적으로 운영할 수 있는 방안을 제시한다. 앞으로는 다양한 LLM 모델을 포함한 시스템 확장과 최적화를 통해 더 많은 사용자 요구를 반영할 수 있을 것으로 기대된다.

참고문헌

[1] ollama github README[웹사이트]. (2024.10.14.). URL:https://github.com/ollama/ollama/blob/main/README.md

[2] Kubernetes Documentation[웹사이트]. (2024.10.14.). URL:https://kubernetes.io/docs/home