vLLM 기본 개념 정리

> - Welcome to vLLM! — vLLM
> - Installation — vLLM
> - vLLM: Easy, Fast, and Cheap LLM Serving with PagedAttention | vLLM Blog
> - 최대 24배 빠른 vLLM의 비밀 파헤치기 – 스캐터랩 기술 블로그 (scatterlab.co.kr)

vLLM?

- LLM Model 은 Model 만 존재해선 사용할 수 없고, 사용해줄 수 있는 Library 가 필요!
  - 기존 이 Library 중 대표적인 친구가 Hugging Face Transformer 이다!
- 이번에 새로 더 빠르게 text 를 생성하는 Library 가 나왔는데 이게 vLLM 이다!
- **즉, vLLM 은 LLM Model 을 사용할 수 있도록 해주는 Library!**
  - PagedAttention 등의 새로운 기술을 사용하여 빠르다! 그 기술적 원리 또한 나열되어 있기는 하나, 이는 사용하지 않겠다!

Docs 정리

Get Started

- vLLM 을 python library 로서, pip install 등을 통해 다운로드 받을 수 있다!
- CUDA 연산을 사용하므로, NVIDIA 환경이 갖춰져야 한다. 자세한 내용은 Installation 부분 살필 것! 또한, AMD Graphic Card 도 지원하기는 하나, 다른 Setup 을 거쳐야 한다!

Quick Start

- Offline Batched Inference : 오프라인에서 추론하기.
- Build API Server : API Server 로 바로 배포하는 것 또한 지원함!
  - host 나 port 에 대한 argument 입력하여 이를 연결할 수 있도록 하는 듯!
- chat completion 등, chat GPT 의 API 와도 호환되어 있다!

> 이후 Distributed Inference and Server, Running on Clouds with skypiolt 등이 있는데, 딱히 할 필요는 없을 듯! => 애초에 나는 Model 에 대한 따로 세부적인 걸 하는 것이 아닌 그냥 Serving 만 하면 되기 때문에, serving tutorial 이나 바로 살펴보고 가장 가능성 있는 것을 시도하는 것이 좋을 듯!