로컬 LLM과 파인튜닝

정규 수업 20-2차시

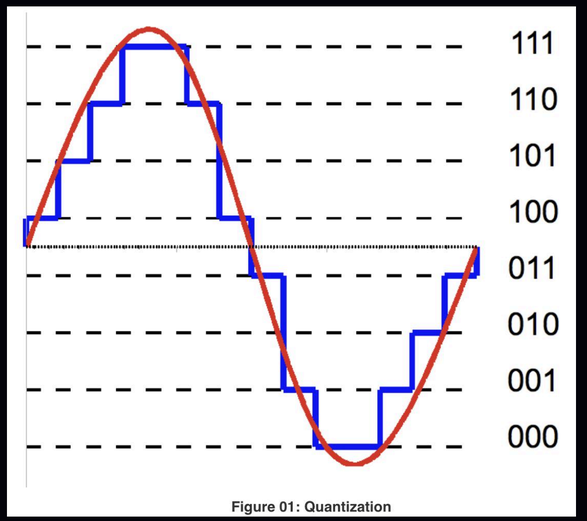
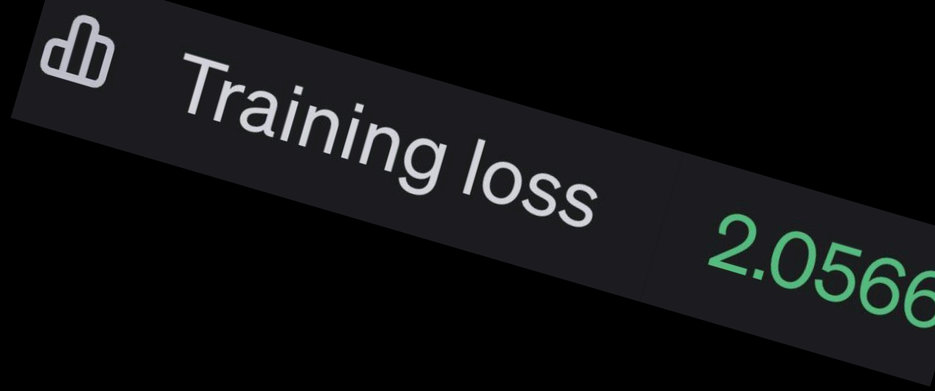


Figure 01: Quantization

LLM API의 한계점



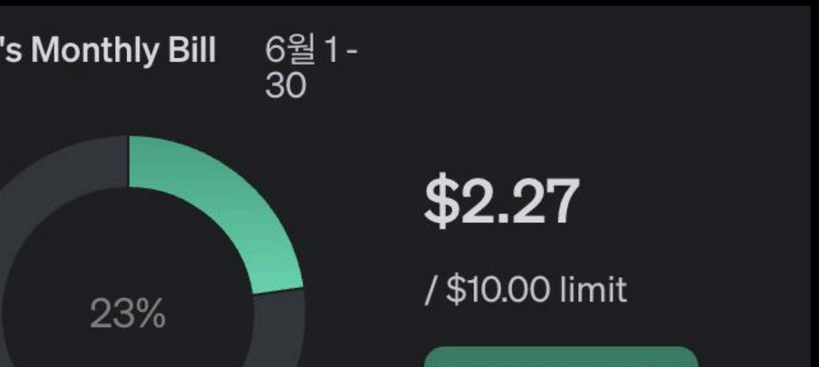
아주 옛날 옛적에...

OpenAI API를 활용해서 GPT를 파인튜닝 했었죠..

??인 것이에요!!!

이러한 LLM API에겐 치명적인 한계점이 있습니다

1. 호출 할 때 마다, 돈을 지불해야한다.
2. 각종 정보(프로필, 대화내용, 모델)가 타인의 서버에 귀속된다.



LLM API의 한계점

호에엥 이걸 어떻게 해결해야하지...?

아하@!!!!!

그냥 파인튜닝도 내 컴퓨터에서!

추론도 내 컴퓨터에서!!

이것을

모델을 로컬(Local)환경에서 구동한다고 합니다.

로컬환경에서 LLM을 구동해보자

헉!! 그거 그래픽카드 엄청 좋아야하는 거 아닌가요?

그래서 우리는 모델을

양자화(quantization)

하기로 했습니다

양자화(quantization)

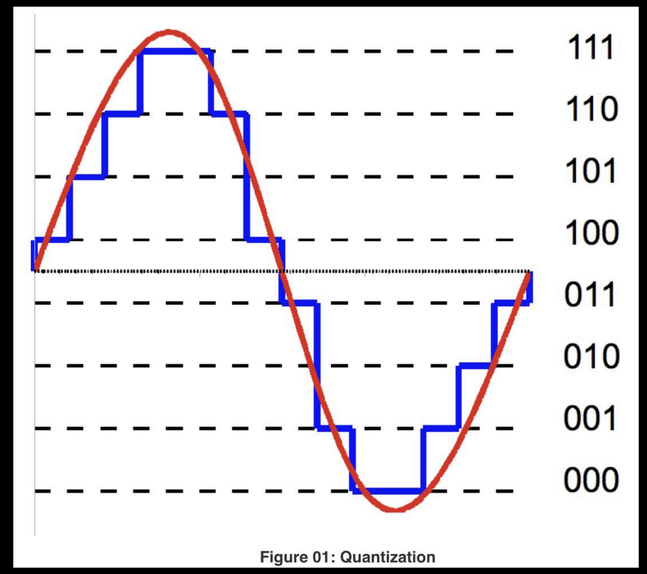


Figure 01: Quantization

모델의 파리미터(가중치 등)를

더 작은 형식으로 변환하여 모델의 연산 효율을 높이는 방법

쉽게 보면 아날로그에 가까운 데이터를 디지털에 가깝게

표현하는 것이라고 설명할 수도 있다

어느 쪽이 계산하기 쉽고 빠를까요?

32비트 부동 소수점 VS 8비트 정수

어느 쪽이 계산하기 쉽고 빠를까요?

32비트 부동 소수점 VS 8비트 정수

그냥 그럼 1bit로 양자화 하면 되는거 아닌가요;;

되겠냐?

양자화는 모델 파라미터의 정보가 일부 손실 되므로

필요 이상의 양자화는 모델 성능에 크게 영향을 줍니다

모델이 거대할 수록 양자화에 있어서 성능 하락의 폭이 작습니다.

2bit, 4bit, 5bit, 8bit, 16bit

등등... 으로 양자화할 수 있습니다.

1bit 양자화는 가능은 하지만... 모델이 모델이 아니지 않을까요..?

참고로 1.58bit, 즉! -1, 0, 1만 표현 가능한

3진법? 양자화는 연구중입니다.

모델을 구동은 한다고 하지만...

파인튜닝은 너무 비용이 비쌉니다.

파인튜닝에 대하여

이러한 부분을 해결하기 위해서
수많은 방법들이 있고 지금도 연구중입니다

파인튜닝에 대해서 조금 더 자세히 알아봅시다!

파인튜닝의 대표적인 방법은 아래와 같습니다

- FFT(Full Fine-Tuning)
- PEFT(Parameter-Efficient Fine-Tuning)

FFT(Full Fine-Tuning)

모델의 모든 파라미터를 조정하는 파인튜닝 방법

비용(돈, 시간)이 어마어마하게 많이 들지만 높은 성능을 기대할 수 있음

PEFT(Parameter-Efficient Fine-Tuning)

모델의 일부 파라미터만 조정하는 파인튜닝 방법

적은 비용으로 원하는 수준의 결과를 만들 수 있음, 각광 받고 있는 방법

오늘은 PEFT 기법을 활용해서 파인튜닝을 할겁니다.

무료 코랩에서 무리 없이 잘 돌아가는 수준으로 작동합니다!

오늘은 PEFT 기법 중에서,

LoRA라는 것을 활용할 것입니다.

LoRA를..

어려운 개념이지만 간단하게 설명하자면...

LoRA(Low-Rank Adaptation)

인공지능 모델은 행렬로서 표현할 수 있는데, 이러한 행렬을 인수분해하여
곱으로 표현함으로서 더 적은 원소를 가지는 행렬로 표현하는 기법

이 방법을 통해서 매우 적은 양의 파라미터를 수정하는 것만으로도
수준 높은 파인튜닝이 가능함