교육
과정
소개서.

LLM 모델 파인튜닝을 위한 Quantization

Fast campus

강의정보

강의장	온라인 강의	데스크탑, 노트북, 모바일 등
수강 기간	평생 소장
상세페이지	https://fastcampus.co.kr/data_online_quantization
강의시간	22시간 31분
문의	고객센터

강의특징

나만의 속도로	낮이나 새벽이나 내가 원하는 시간대에 나의 스케쥴대로 수강
원하는 곳 어디서나	시간을 쪼개 먼 거리를 오가며 오프라인 강의장을 찾을 필요 없이 어디서든 수강
무제한 복습	무엇이든 반복적으로 학습해야 내것이 되기에 이해가 안가는 구간 몇번이고 재생

Copyright ⓒ FAST CAMPUS Corp. All Rights Reserved.

강의목표

- LLM의 파라미터와 계산 복잡도가 모델 크기와 성능에 따라 어떻게 영향을 미치는지 알아보고, LLM의 메모리 절약과 연산 속도를 올리기 위한 Quantization의 기초 개념을 학습합니다.
- LLM 모델을 Quantization 할 때 활용되는 대표적인 2가지 방법 PTQ, QAT의 개념을 학습합니다.
- LLM 모델을 훈련 후 양자화하는 방법인 PTQ 방법론의 개념을 이해하고, PTQ 방법에서 활용하는 대표 3가지 방법을 실습합니다.
- 기존 Quantization 진행 중 정수 연산에서 Latency 발생과 저비트로 변환하면서 생기는 Outliar와 오차를 막기 위한 최신 Quantization 기법들을 학습합니다.

강의요약

- RTX 3090만으로도 NVIDIA A100 수준의 컴퓨팅 파워를 낼 수 있는 방법, 이승유 강사님의 LLM Quantization 노하우를 바로 확인하세요!
- 패키지 상품을 선택하고 LLM 파인 튜닝에 대한 개념부터 실전 LLM 서비스 개발을 위한 노하우까지 평생 소장하세요.
- 극한의 GPU 메모리 환경에서 Domain Specific LLM 모델을 개발하기 위한 합성 데이터 생성부터 SFT & QLoRA Inference 프로젝트 실습을 직접 구현합니다.
- LLM 모델을 Quantization 할 때 활용되는 대표적인 2가지 방법 PTQ, QAT의 개념을 학습합니다.

Copyright ⓒ FAST CAMPUS Corp. All Rights Reserved.

강사

이승유

과목	-	LLM 모델 파인튜닝을 위한 Quantization
약력	-	현 Markr AI LLM Researcher
	-	GPU 최적화와 Fine-Tuning 기술로 자체 LLM 모델 개발하여 Open-Ko LLM & Open LLM Leaderboard 최장기간 1위 달성
	-	PEFT(Parameter Efficient Fine-Tuning) 라이브러리 Contributor

Copyright ⓒ FAST CAMPUS Corp. All Rights Reserved.

CURRICULUM

01.

LLM과 Quantization 기초

파트별 수강시간 05:05:05

CH01. 강의 소개
01. 강의 흐름 및 강의 소개
02. 학습 목표 및 기대효과
CH02. LLM과 Quantization의 배경과 연관성
01. LLM 구조 및 동작 방식
02. LLM 경량화 방법론들 소개
03. Quantization의 등장 배경
CH03. Quantization의 기초 개념
01. Quantization의 개념
02. 양자화의 장단점 비교
03. 양자화 접근 방법 - PTQ, QAT
04. 다양한 양자화 기법들 소개
CH04. Quantization의 최신 트랜드
01. Qunatization의 최신 트렌드

CURRICULUM

02.

Fine-Tuning을 위한 Quantization

파트별 수강시간 04:49:03

CH01. Quantization-Aware-Training(QAT)
01. QAT의 기초 개념
02. QAT 방법론의 장단점 및 활용
CH02. QLoRA를 활용한 Fine-Tuning
01. 다양한 Fine-tuning 방법론들
02. QLoRA 방법론 소개
03. QLoRA 심화 내용
CH03. [실습] QLoRA를 활용한 Fine-tuning
01. QLoRA 활용 Fine-tuning 실습
02. 결과 분석 및 최적화 방법론 논의
03. 결과 분석 (Transformers 라이브러리 이슈)

CURRICULUM

03.

훈련 후 Quantization 방법론

파트별 수강시간 05:41:08

CH01. Post-Training Quantization(PTQ)
01. PTQ의 기초 개념
CH02. GGUF Quantization
01. Llama.cpp 라이브러리 설명
02. [실습] LLaMA.cpp 라이브러리를 활용한 Quantization 실습
CH03. GPT-Q Quantization
01. GPT-Q의 개념 설명
02. [실습] GPT-Q를 활용한 Quantization
CH04. AWQ Quantization
01. AWQ의 개념 설명
02. [실습] AWQ를 활용한 Quantization
CH05. Quantization 정리
01. Quantization 방법론 정리 (Recap)

CURRICULUM

04.

양자화 기법 심화

파트별 수강시간 04:21:13

CH01. 고급 양자화 기법
01. 최신 양자화 기법
02. 양자화 방법론 논의 및 심화
CH02. QLoRA를 활용한 다양한 Fine-tuning Tip
01. QLoRA로 훈련시킬 시 주의사항
02. Hyperparameter 분석 및 설명
03. Unsloth 라이브러리 설명
04. [실습] Unsloth를 활용한 Fine-Tuning
05. Fine-Tuning의 장단점

CURRICULUM

05.

프로젝트 실습 및 응용

파트별 수강시간 02:34:32

CH01. 실제 프로젝트 기반 FineTuning 실습
01. 데이터 준비 및 전처리
02. QLoRA 기반 SFT 방법론 실습
03. 실제 프로젝트 기반 PTQ 방법론 실습 및 적용
CH02. Inference 실습
01. PTQ - Quantized Model Inference 실습
02. Inference 실습 심화

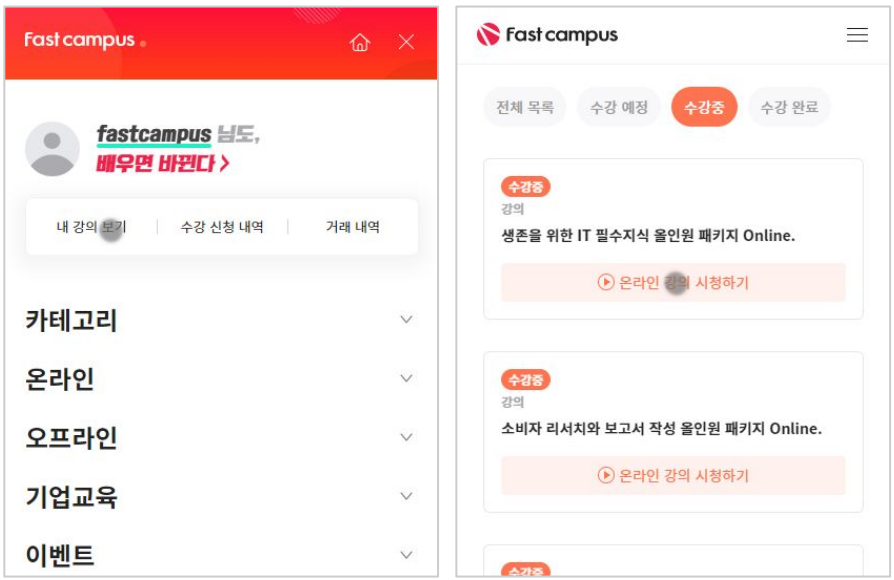
본 과정은 현재 촬영 및 편집이 진행되고 있는 사전 판매 중인 강의입니다.
해당 교육과정 소개서는 변경되거나 추가될 수 있습니다.

주의 사항

- 상황에 따라 사전 공지 없이 할인이 조기 마감되거나 연장될 수 있습니다.
- 패스트캠퍼스의 모든 온라인 강의는 **아이디 공유를 금지**하고 있으며 1개의 아이디로 여러 명이 수강하실 수 없습니다.
- 별도의 주의사항은 각 강의 상세페이지에서 확인하실 수 있습니다.

수강 방법

- 패스트캠퍼스는 크롬 브라우저에 최적화 되어있습니다.
- 사전 예약 판매 중인 강의의 경우 1차 공개일정에 맞춰 '온라인 강의 시청하기'가 활성화됩니다.



환불 규정

- 온라인 강의는 각 과정 별 '정상 수강기간(유료수강기간)'과 정상 수강기간 이후의 '복습 수강기간(무료수강기간)'으로 구성됩니다.
- 환불금액은 실제 결제금액을 기준으로 계산됩니다.

수강 시작 후 7일 이내	100% 환불 가능 (단, 수강하셨다면 수강 분량만큼 차감)
수강 시작 후 7일 경과	정상(유료) 수강기간 대비 잔여일에 대해 환불규정에 따라 환불 가능

※ 강의별 환불규정이 상이할 수 있으므로 각 강의 상세페이지를 확인해 주세요.

Copyright ⓒ FAST CAMPUS Corp. All Rights Reserved.