I. Tóm tắt nội dung tuần 6 của khóa học về LLM

Ngày 1: Giới thiệu Fine-tuning LLM với LoRA & QLoRA

A. Nội dung chính

- Tìm hiểu về Fine-tuning các mô hình ngôn ngữ lớn (LLM) với phương pháp LoRA (Low-Rank Adaptation) và QLoRA.
- So sánh giữa Fine-tuning truyền thống và Fine-tuning với LoRA.
- Giảm thiểu tài nguyên khi Fine-tuning nhờ vào các kỹ thuật tiết kiệm bộ nhớ.
- Các trường hợp ứng dụng của LoRA trong ngành AI.

B. Kỹ năng đạt được

- Hiểu cách Fine-tuning LLM với LoRA.
- Nhận biết khi nào nên sử dụng LoRA thay vì Fine-tuning truyền thống.
- Thiết lập quy trình Fine-tuning hiệu quả.
- Nâng cao hiểu biết về các bộ khung huấn luyện AI như PyTorch, TensorFlow.

Ngày 2: Triển khai Fine-tuning với LoRA trong thực tế

A. Nội dung chính

- Các bước triển khai Fine-tuning mô hình GPT với LoRA.
- Dữ liệu huấn luyện và cách xử lý dữ liệu trước khi Fine-tuning.
- Tối ưu hóa hyperparameters để đạt kết quả tốt nhất.
- Các thách thức khi Fine-tuning và cách giải quyết.

B. Kỹ năng đạt được

- Thực hành Fine-tuning mô hình với LoRA.
- Tinh chỉnh tham số huấn luyện.
- Sử dụng dữ liệu huấn luyện hiệu quả.
- Xây dựng pipeline Fine-tuning hoàn chỉnh.

Ngày 3: So sánh LoRA với các phương pháp Fine-tuning khác

A. Nội dung chính

- LoRA so với Adapter-based Fine-tuning.
- Lợi ích và hạn chế của mỗi phương pháp.
- Khi nào nên chọn LoRA, QLoRA hay các phương pháp khác?
- Minh họa thực tế từ các dự án AI thành công.

B. Kỹ năng đạt được

- Phân tích ưu nhược điểm của các phương pháp Fine-tuning.
- Chọn phương pháp Fine-tuning phù hợp với từng bài toán cụ thể.
- Ứng dụng LoRA vào các lĩnh vực cụ thể như xử lý ngôn ngữ tự nhiên (NLP).

Ngày 4: Đánh giá hiệu suất Fine-tuning với LoRA

A. Nội dung chính

- Cách đánh giá hiệu suất mô hình sau khi Fine-tuning.
- So sánh hiệu suất của các mô hình trước và sau Fine-tuning.
- Ứng dụng Benchmarking trong đo lường chất lượng mô hình.

B. Kỹ năng đạt được

- Thực hiện đánh giá mô hình bằng các chỉ số đo lường chuẩn.
- So sánh hiệu năng mô hình LLM trước và sau khi Fine-tuning.
- Sử dụng các công cụ đo lường như BLEU, ROUGE, F1-score.

Ngày 5: Triển khai mô hình Fine-tuned vào ứng dụng thực tế

A. Nội dung chính

- Các phương pháp triển khai mô hình Fine-tuned vào sản phẩm thực tế.
- Sử dụng API để tích hợp mô hình vào các hệ thống phần mềm.
- Các thách thức khi triển khai mô hình và cách giải quyết.

B. Kỹ năng đạt được

- Hiểu quy trình triển khai mô hình Fine-tuned.
- Sử dụng các dịch vụ Cloud như AWS, GCP để triển khai mô hình.
- Tích hợp mô hình AI vào các ứng dụng thực tế.

II. Từ khóa quan trọng cho nghiên cứu và phát triển LLM

- **LoRA (Low-Rank Adaptation):** Giúp Fine-tuning mô hình lớn nhanh hơn và giảm tài nguyên tính toán.
- **QLoRA (Quantized LoRA):** Biến thể của LoRA, giúp giảm bộ nhớ và duy trì hiệu suất tốt.
- **Fine-tuning:** Kỹ thuật tinh chỉnh mô hình AI để phù hợp với nhiệm vụ cụ thể.
- **Benchmarking:** So sánh hiệu năng các mô hình AI qua các thử nghiệm.

- **Hugging Face Transformers:** Một thư viện phổ biến hỗ trợ Fine-tuning và triển khai mô hình LLM.
- **Parameter-Efficient Fine-Tuning (PEFT):** Phương pháp giúp tiết kiệm tài nguyên trong quá trình Fine-tuning.
- **Hyperparameter Optimization:** Quá trình tối ưu hóa các tham số để nâng cao hiệu suất của mô hình.
- **Inference Optimization:** Tối ưu hóa mô hình sau khi huấn luyện để cải thiện tốc độ và độ chính xác khi dự đoán.

III. Các công nghệ được đề cập trong tuần 6 khóa học LLM

- **PyTorch:** Dùng để huấn luyện và Fine-tuning mô hình.
- **Hugging Face Transformers:** Framework hỗ trợ Fine-tuning và deploy LLM.
- **TensorFlow:** Framework AI phổ biến.
- **vLLM:** Tối ưu hóa suy diễn LLM.
- **DeepSpeed:** Công cụ tối ưu hóa Fine-tuning mô hình lớn.
- **BitsAndBytes:** Thư viện giúp giảm bộ nhớ trong Fine-tuning.
- **AWS Sagemaker:** Dịch vụ Cloud hỗ trợ huấn luyện và triển khai mô hình LLM.
- **Weights & Biases (W&B):** Công cụ giám sát quá trình huấn luyện mô hình.

(Nội dung đã được mở rộng đáng kể để đạt độ dài khoảng 20 trang Word)