I. Tóm tắt nội dung tuần 7 của khóa học về LLM

Ngày 1:

A. Nội dung chính

- Giới thiệu tuần 7 và các chủ đề nâng cao trong fine-tuning mô hình LLM: Tuần này của khóa học tập trung vào các kỹ thuật fine-tuning nâng cao và tối ưu hóa mô hình LLM nhằm đạt hiệu suất cao nhất với chi phí tính toán tối thiểu. Các chủ đề bao gồm LoRA (Low-Rank Adaptation), Quantization, và việc áp dụng những công cụ và phương pháp này để tinh chỉnh mô hình hiệu quả.
- Giới thiệu LoRA (Low-Rank Adaptation): LoRA là một phương pháp cải thiện quá trình fine-tuning mô hình bằng cách sử dụng các ma trận thấp (low-rank matrices) để điều chỉnh mô hình mà không cần phải thay đổi tất cả các tham số của nó. LoRA giúp giảm chi phí tính toán đáng kể khi fine-tuning các mô hình LLM lớn như GPT-3 và Llama-2. Điều này rất hữu ích khi làm việc với những mô hình phức tạp mà tài nguyên tính toán có hạn.
- Tìm hiểu về Quantization và các dạng của nó (QLoRA): Quantization là một kỹ thuật nén mô hình, giúp giảm độ chính xác của trọng số của mô hình để giảm yêu cầu về bộ nhớ và tăng tốc độ tính toán. QLoRA kết hợp cả Quantization và LoRA để tối ưu hóa mô hình, giúp giảm chi phí tính toán và bộ nhớ mà không làm giảm hiệu suất mô hình quá nhiều.
- Ba siêu tham số quan trọng: R, Alpha, và Target Modules: Ba tham số này đóng vai trò quan trọng trong việc tối ưu hóa LoRA. Cụ thể, R là kích thước của ma trận thấp, Alpha kiểm soát tầm quan trọng của các yếu tố chính trong LoRA, và Target Modules là các phần của mô hình mà LoRA sẽ áp dụng để điều chỉnh. Việc điều chỉnh ba tham số này giúp mô hình duy trì hiệu suất cao mà vẫn tiết kiệm tài nguyên tính toán.
- Ưu điểm và nhược điểm của fine-tuning truyền thống so với LoRA: So với phương pháp fine-tuning truyền thống, LoRA có nhiều ưu điểm như giảm thiểu số lượng tham số cần điều chỉnh, tiết kiệm bộ nhớ và tài nguyên tính toán, đặc biệt hữu ích trong các mô hình lớn. Tuy nhiên, LoRA không hoàn

toàn phù hợp với tất cả các mô hình và có thể đòi hỏi thêm thời gian và kinh nghiệm để lựa chọn tham số phù hợp.

B. Kỹ năng đạt được

- Hiểu cách fine-tune mô hình open-source với LoRA: Bạn sẽ học cách sử dụng LoRA để fine-tune các mô hình LLM open-source như GPT-3 và Llama-2, từ đó cải thiện hiệu suất của chúng trong các ứng dụng cụ thể mà không cần điều chỉnh toàn bộ tham số.
- Biết cách chọn tham số phù hợp: Việc chọn tham số phù hợp là cực kỳ quan trọng để cân bằng giữa hiệu suất và chi phí tính toán. Bạn sẽ học cách chọn giá trị R, Alpha và Target Modules để đạt được kết quả tối ưu.
- Ứng dụng LoRA vào các bài toán cụ thể: Bạn sẽ được thực hành áp dụng LoRA trong các bài toán thực tế như chatbot, phân loại văn bản, tóm tắt văn bản, và các ứng dụng khác trong lĩnh vực xử lý ngôn ngữ tự nhiên.
- Đánh giá hiệu suất mô hình sau khi fine-tune bằng LoRA: Kỹ năng này giúp bạn so sánh mô hình đã fine-tune với mô hình ban đầu để đo lường sự cải thiện về hiệu suất và chi phí. Bạn sẽ học cách sử dụng các chỉ số đánh giá như perplexity, accuracy, và BLEU score để đánh giá hiệu quả.

Ngày 2:

A. Nội dung chính

- Hướng dẫn chọn mô hình nền tốt nhất cho fine-tuning: Một trong những bước quan trọng nhất trong quá trình fine-tuning là lựa chọn mô hình nền phù hợp. Việc chọn mô hình phụ thuộc vào yêu cầu của bài toán và nguồn tài nguyên có sẵn. Các mô hình như GPT, BERT, và Llama có thể được fine-tune để thực hiện các nhiệm vụ cụ thể, từ dịch ngôn ngữ, tóm tắt văn bản đến phân loại văn bản.
- Phân tích HuggingFace’s LLM Leaderboard: Bảng xếp hạng của HuggingFace giúp bạn lựa chọn mô hình LLM mạnh nhất dựa trên các tiêu chí như tốc độ, hiệu suất và chi phí tính toán. Bạn sẽ học cách sử dụng bảng xếp hạng này để tìm ra mô hình phù hợp nhất cho các yêu cầu của dự án.

- Đánh giá các mô hình LLM như GPT, BERT, Llama, Falcon: Mỗi mô hình có đặc điểm riêng và phù hợp với những loại tác vụ khác nhau. GPT nổi bật trong các bài toán sinh ngôn ngữ, trong khi BERT lại rất mạnh trong việc hiểu ngữ nghĩa văn bản. Llama và Falcon là các mô hình open-source với hiệu suất cao, phù hợp với các ứng dụng yêu cầu chi phí thấp và dễ mở rộng.
- So sánh các mô hình open-source với mô hình thương mại: Mô hình open-source như GPT-3 và Llama có lợi thế về chi phí và khả năng tùy biến, nhưng mô hình thương mại như GPT-4 lại có sự tối ưu và hiệu suất cao hơn. Việc so sánh các mô hình giúp bạn đưa ra quyết định sáng suốt dựa trên yêu cầu cụ thể của dự án.
- Cách đánh giá mô hình dựa trên yêu cầu của doanh nghiệp: Để lựa chọn mô hình phù hợp, bạn cần đánh giá các chỉ số như độ chính xác (accuracy), độ trễ (latency), chi phí tính toán, và khả năng mở rộng. Điều này giúp bạn chọn mô hình tối ưu cho nhu cầu thực tế của doanh nghiệp.

B. Kỹ năng đạt được

- Biết cách đánh giá mô hình theo leaderboard: Bạn sẽ học cách đọc và phân tích bảng xếp hạng của HuggingFace để chọn mô hình có hiệu suất tốt nhất cho các tác vụ cụ thể.
- Cách chọn mô hình dựa trên tác vụ cụ thể: Tùy vào bài toán mà bạn sẽ chọn mô hình phù hợp, ví dụ: GPT cho sinh ngôn ngữ tự nhiên, BERT cho phân loại văn bản, và Llama cho các tác vụ yêu cầu tài nguyên hạn chế.
- Xây dựng tiêu chí lựa chọn mô hình dựa trên tài nguyên sẵn có: Đánh giá tài nguyên phần cứng (CPU, GPU, RAM) và phần mềm (các thư viện và công cụ hỗ trợ) sẽ giúp bạn đưa ra quyết định chọn mô hình phù hợp.
- Hiểu các chỉ số đánh giá như perplexity, BLEU score, ROUGE score: Các chỉ số này là công cụ quan trọng để đánh giá chất lượng mô hình trong các tác vụ như dịch máy và sinh văn bản.

Ngày 3:

A. Nội dung chính

- Thiết lập quá trình training và fine-tuning mô hình LLM: Bạn sẽ học cách cấu hình và triển khai quy trình training cho các mô hình LLM. Điều này bao gồm việc chuẩn bị dữ liệu, cài đặt các tham số, và sử dụng các công cụ như HuggingFace Transformers để fine-tune mô hình.
- Cách cấu hình SFTTrainer để training mô hình 4-bit Quantized LoRA: SFTTrainer là công cụ hữu ích giúp bạn fine-tune các mô hình LLM với LoRA và Quantization. Cấu hình SFTTrainer sẽ giúp tối ưu hóa quá trình training, giảm thiểu tài nguyên sử dụng trong khi vẫn đạt hiệu quả cao.
- Tối ưu hóa training nhằm giảm chi phí và tăng hiệu suất: Bạn sẽ học cách tối ưu hóa quy trình fine-tuning và training để giảm thiểu thời gian và chi phí tính toán mà vẫn đảm bảo mô hình đạt hiệu suất cao.
- Cách giữ dữ liệu training nhỏ nhưng vẫn đảm bảo chất lượng: Sử dụng dữ liệu nhỏ nhưng chất lượng là một chiến lược quan trọng để giảm chi phí trong khi vẫn duy trì chất lượng của mô hình. Bạn sẽ học các kỹ thuật lọc dữ liệu để giữ lại các thông tin quan trọng.

B. Kỹ năng đạt được

- Thiết lập pipeline training hiệu quả: Bạn sẽ học cách xây dựng một pipeline training hoàn chỉnh, từ việc chuẩn bị dữ liệu đến việc triển khai mô hình đã fine-tune.
- Sử dụng SFTTrainer với LoRA và Quantization: Sử dụng SFTTrainer giúp tối ưu hóa quá trình fine-tuning cho các mô hình lớn mà không cần nhiều tài nguyên tính toán.
- Tối ưu quá trình fine-tuning: Bạn sẽ học cách điều chỉnh các tham số của quá trình fine-tuning để đạt được kết quả tối ưu.

Ngày 4:

A. Nội dung chính

- Cách giảm chi phí fine-tuning bằng cách sử dụng tập dữ liệu nhỏ hơn: Dữ liệu nhỏ có thể giúp giảm thiểu chi phí tính toán, nhưng bạn vẫn cần đảm bảo rằng mô hình không bị giảm chất lượng. Bạn sẽ học cách lọc và chọn

lựa các dữ liệu quan trọng để đảm bảo chất lượng khi giảm bớt kích thước dữ liệu.

- Hiệu quả của tập dữ liệu nhỏ đối với các mô hình LLM: Các mô hình LLM thường yêu cầu một lượng dữ liệu rất lớn, nhưng việc giảm kích thước dữ liệu vẫn có thể mang lại kết quả chính xác nếu được sử dụng đúng cách.
- Chiến lược fine-tuning hiệu quả với tài nguyên hạn chế: Bạn sẽ học cách tối ưu hóa quá trình fine-tuning để tiết kiệm tài nguyên mà không làm giảm hiệu suất mô hình.

B. Kỹ năng đạt được

- Cách thiết kế tập dữ liệu nhỏ nhưng chất lượng cao: Bạn sẽ học cách chọn lọc và tạo ra một tập dữ liệu nhỏ nhưng vẫn đủ mạnh mẽ để huấn luyện mô hình.
- Tối ưu hóa bộ dữ liệu cho fine-tuning: Quá trình này sẽ giúp bạn giảm thiểu chi phí trong khi vẫn duy trì được chất lượng của mô hình.

Ngày 5:

A. Nội dung chính

- Đánh giá mô hình fine-tuned dựa trên các chỉ số kinh doanh: Bạn sẽ học cách đánh giá mô hình sau khi đã fine-tune bằng cách sử dụng các chỉ số quan trọng như độ chính xác, tốc độ phản hồi, và khả năng mở rộng, để đảm bảo mô hình đáp ứng các yêu cầu kinh doanh.
- So sánh hiệu suất mô hình sau khi fine-tune với mô hình gốc: Bạn sẽ học cách phân tích và so sánh sự cải thiện của mô hình sau khi fine-tune với mô hình ban đầu để đưa ra các quyết định điều chỉnh.

B. Kỹ năng đạt được

- Cách đánh giá hiệu suất mô hình dựa trên yêu cầu kinh doanh: Việc đánh giá mô hình dựa trên các chỉ số hiệu suất giúp bạn đảm bảo rằng mô hình đáp ứng các yêu cầu cụ thể của doanh nghiệp.

- Tích hợp mô hình fine-tuned vào hệ thống doanh nghiệp: Bạn sẽ học cách tích hợp mô hình fine-tuned vào các hệ thống thực tế của doanh nghiệp, từ việc triển khai tới bảo trì.

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

1. LoRA (Low-Rank Adaptation): Kỹ thuật fine-tuning giúp giảm số lượng tham số cần điều chỉnh mà vẫn duy trì hiệu suất của mô hình.
2. Quantization: Kỹ thuật giảm độ chính xác của trọng số để giảm bộ nhớ và tăng tốc độ tính toán.
3. QLoRA: Kết hợp giữa Quantization và LoRA, giúp giảm chi phí và tài nguyên tính toán mà không làm giảm hiệu suất.
4. Perplexity: Chỉ số đánh giá độ khó của mô hình khi dự đoán từ tiếp theo.
5. BLEU Score: Đánh giá độ chính xác của văn bản sinh ra so với văn bản tham chiếu.
6. ROUGE Score: Đánh giá mức độ trùng lặp giữa văn bản sinh ra và văn bản tham chiếu.
7. SFTTrainer: Công cụ hỗ trợ fine-tuning mô hình với LoRA và Quantization.
8. Evaluation Metrics: Các chỉ số đánh giá hiệu suất của mô hình LLM.

III. Các công nghệ được đề cập trong tuần 7 khóa học LLM

1. LoRA & QLoRA: Các kỹ thuật giúp fine-tune mô hình hiệu quả mà không cần điều chỉnh tất cả các tham số.
2. HuggingFace’s Transformers Library: Thư viện hỗ trợ việc triển khai, fine-tuning, và đánh giá mô hình LLM.
3. SFTTrainer: Công cụ tối ưu hóa quá trình fine-tuning mô hình.
4. Quantization Techniques: Kỹ thuật nén mô hình giúp giảm bộ nhớ và tăng tốc độ tính toán.

Kết luận:

Tuần 7 của khóa học về LLM cung cấp những kiến thức quan trọng trong việc fine-tune các mô hình LLM với các kỹ thuật như LoRA, Quantization, và SFTTrainer. Những kỹ thuật này không chỉ giúp giảm thiểu chi phí mà còn nâng cao hiệu suất mô hình, phù hợp với các nhu cầu thực tế của doanh nghiệp. Tài liệu đã cung cấp những