I. Tóm tắt nội dung Week 03 của khóa học về LLM

Ngày 1: Giới thiệu tổng quan về Hugging Face, cách sử dụng Hugging Face Hub và Google Colab.

A. Nội dung chính:

- Giới thiệu tổng quan về Hugging Face, một nền tảng phổ biến trong cộng đồng AI.
- Khám phá Hugging Face Hub: cách tìm kiếm, tải về và sử dụng các mô hình AI có sẵn.
- Giới thiệu về Google Colab, cách thiết lập môi trường làm việc và chạy mã Python trên nền tảng này.
- Hướng dẫn cách sử dụng GPU miễn phí trên Google Colab để tăng tốc xử lý mô hình AI.

B. Kỹ năng đạt được:

- Hiểu rõ vai trò và lợi ích của Hugging Face trong phát triển mô hình AI.
- Thành thạo cách truy cập và tìm kiếm mô hình AI trên Hugging Face Hub.
- Thiết lập và chạy môi trường lập trình AI trên Google Colab.
- Sử dụng GPU trên Colab để chạy mô hình AI hiệu quả hơn.

Ngày 2: Làm quen với Hugging Face Pipelines, thực hành phân tích cảm xúc, nhận diện thực thể, tóm tắt văn bản và dịch ngôn ngữ

A. Nội dung chính:

- Giới thiệu Hugging Face Pipelines: một API cấp cao giúp đơn giản hóa việc sử dụng mô hình AI.
- Hướng dẫn cách chạy các pipeline phổ biến như:
    - Phân tích cảm xúc (Sentiment Analysis).
    - Nhận diện thực thể có tên (Named Entity Recognition - NER).
    - Tóm tắt văn bản (Text Summarization).
    - Dịch ngôn ngữ (Translation).
- Thực hành với Google Colab để chạy các pipeline trên dữ liệu thực tế.

B. Kỹ năng đạt được:

- Hiểu rõ về Pipelines API và ứng dụng của nó trong xử lý ngôn ngữ tự nhiên.
- Sử dụng các pipeline AI để thực hiện các tác vụ khác nhau mà không cần huấn luyện lại mô hình.
- Tích hợp pipeline vào ứng dụng thực tế bằng Python.

Ngày 3: Tìm hiểu về Tokenizers, cách chuyển đổi văn bản thành token, và thực hành với các mô hình LLama, Phi-2, Qwen, Starcoder.

A. Nội dung chính:

- Tìm hiểu sâu về Tokenizers trong Hugging Face: cách chuyển đổi văn bản thành token.
- Các loại Tokenizers phổ biến và cách chúng hoạt động.
- Khám phá các mô hình ngôn ngữ mã nguồn mở như:
    - LLama (Meta AI)
    - Phi-2 (Microsoft)
    - Qwen (Alibaba Cloud)
    - Starcoder (dành cho lập trình viên)
- Thực hành tạo và sử dụng tokenizers trên các mô hình AI.

B. Kỹ năng đạt được:

- Hiểu cách tokenization ảnh hưởng đến hiệu suất của mô hình AI.
- Biết cách chọn tokenizer phù hợp với từng loại mô hình.
- Sử dụng tokenizers để xử lý dữ liệu đầu vào một cách hiệu quả.

Ngày 4: Làm việc với Model Class, chạy suy diễn trên mô hình mở và ứng dụng Quantization để tối ưu mô hình.

A. Nội dung chính:

- Tìm hiểu về Model Class trong Hugging Face: cách hoạt động của các mô hình AI.
- Hướng dẫn tải mô hình và chạy suy diễn (inference) với các mô hình mở.
- Giới thiệu về Quantization - kỹ thuật giúp giảm dung lượng mô hình mà vẫn duy trì hiệu suất cao.
- Thực hành chạy inference trên nhiều mô hình khác nhau để so sánh kết quả.

B. Kỹ năng đạt được:

- Hiểu rõ cách hoạt động của Model Class trong Hugging Face.
- Biết cách tải và triển khai mô hình AI từ Hugging Face.
- Sử dụng Quantization để tối ưu mô hình cho các thiết bị có tài nguyên hạn chế.

Ngày 5: Xây dựng hệ thống nhận diện giọng nói và tóm tắt nội dung cuộc họp bằng cách kết hợp OpenAI Whisper và mô hình Hugging Face.

A. Nội dung chính:

- Tích hợp mô hình mở và có phí để xây dựng ứng dụng AI hoàn chỉnh.
- Xây dựng hệ thống chuyển đổi âm thanh thành văn bản và tóm tắt nội dung cuộc họp.
- Sử dụng mô hình OpenAI Whisper để nhận diện giọng nói.
- Dùng mô hình Hugging Face để tóm tắt nội dung và trích xuất thông tin quan trọng.
- Thực hành triển khai mô hình trên Google Colab và kiểm tra kết quả.

B. Kỹ năng đạt được:

- Kết hợp nhiều mô hình AI khác nhau để giải quyết một bài toán phức tạp.
- Sử dụng API OpenAI và Hugging Face một cách thành thạo.
- Xây dựng pipeline AI hoàn chỉnh từ xử lý âm thanh đến phân tích nội dung.

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

2.1.Tokenization

- Chuyển đổi văn bản thành các đơn vị nhỏ hơn (token) để mô hình AI có thể xử lý.
- Các phương pháp phổ biến: Word-based, Subword-based, Character-based.
- Giúp cải thiện độ chính xác của mô hình khi làm việc với các ngôn ngữ có cấu trúc khác nhau.
- Ứng dụng trong NLP: phân loại văn bản, dịch máy, tóm tắt văn bản.

2.2.Model Quantization

- Giảm kích thước mô hình bằng cách sử dụng ít bit hơn để lưu trữ trọng số.
- Các phương pháp phổ biến: 8-bit quantization, 4-bit quantization.
- Giúp giảm mức sử dụng bộ nhớ và tăng tốc độ suy diễn mô hình trên thiết bị phần cứng yếu.
- Ứng dụng trong triển khai AI trên điện thoại di động và thiết bị IoT.

2.3.Hugging Face Pipelines

- API cấp cao giúp đơn giản hóa việc chạy mô hình AI.
- Hỗ trợ nhiều tác vụ: phân tích cảm xúc, nhận diện thực thể, tóm tắt văn bản, dịch thuật.

- Giúp tiết kiệm thời gian phát triển và thử nghiệm mô hình AI.
- Ứng dụng trong việc tạo chatbot, hệ thống phân tích dữ liệu.

2.4.Reinforcement Learning from Human Feedback (RLHF)

- Kỹ thuật học tăng cường từ phản hồi của con người để cải thiện mô hình AI.
- Giúp mô hình AI học được cách phản hồi phù hợp với ngữ cảnh thực tế.
- Ứng dụng trong AI hội thoại, chatbot thông minh, trợ lý ảo.

2.5.Transfer Learning trong AI

- Sử dụng một mô hình đã được huấn luyện trước trên một tập dữ liệu lớn, sau đó tinh chỉnh cho một tác vụ cụ thể.
- Giúp tiết kiệm tài nguyên và thời gian huấn luyện.
- Ứng dụng trong nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên, phân tích y khoa.

2.6.Fine-tuning

- Tinh chỉnh mô hình AI bằng cách đào tạo trên một tập dữ liệu đặc thù.
- Giúp mô hình hoạt động tốt hơn trong lĩnh vực cụ thể.
- Ứng dụng trong xây dựng chatbot doanh nghiệp, phân loại tin tức.

2.7.Open-Source LLMs

- Mô hình ngôn ngữ lớn mã nguồn mở như LLama, Phi-2, Qwen giúp cộng đồng tiếp cận dễ dàng.
- Cho phép nghiên cứu, cải tiến, và tùy chỉnh theo nhu cầu.
- Ứng dụng trong giáo dục, nghiên cứu AI.

2.8.Low-Level APIs

- API cấp thấp giúp lập trình viên kiểm soát chi tiết hơn việc huấn luyện và chạy mô hình.
- Cho phép tối ưu hóa tokenization, kiến trúc mô hình.
- Ứng dụng trong nghiên cứu và phát triển AI nâng cao.

2.9.Self-Supervised Learning

- Phương pháp học máy mà mô hình tự tìm hiểu từ dữ liệu không cần gán nhãn.
- Giúp tiết kiệm chi phí thu thập dữ liệu.
- Ứng dụng trong phát triển chatbot, tìm kiếm thông tin thông minh.

2.10.Bit and Bytes Quantization

- Giảm số bit đại diện cho trọng số mô hình để tối ưu hóa bộ nhớ.
- Phù hợp với triển khai AI trên các thiết bị có tài nguyên hạn chế.
- Ứng dụng trong AI nhúng, thiết bị IoT, di động.

III. Các công nghệ được đề cập trong Week 03

3.1.Hugging Face Transformers

- Thư viện mạnh mẽ hỗ trợ nhiều mô hình AI trong xử lý ngôn ngữ tự nhiên (NLP), bao gồm các mô hình Transformer như BERT, GPT, LLama, và nhiều mô hình khác.
- **Ý nghĩa**: Giúp đơn giản hóa việc triển khai, huấn luyện và tinh chỉnh mô hình ngôn ngữ lớn.
- **Vai trò & Ứng dụng**:
  - Cung cấp giao diện đơn giản để tải và sử dụng các mô hình NLP.
  - Ứng dụng trong chatbot, phân tích ngữ nghĩa, và tạo nội dung tự động.

3.2.OpenAI Whisper

- **Là gì?**: Mô hình nhận diện giọng nói mạnh mẽ từ OpenAI, có khả năng chuyển đổi giọng nói thành văn bản với độ chính xác cao.
- **Ý nghĩa**: Giúp cải thiện khả năng chuyển đổi âm thanh thành văn bản, hỗ trợ nhiều ngôn ngữ.
- **Vai trò & Ứng dụng**:
  - Ứng dụng trong trợ lý ảo, công cụ phiên dịch giọng nói, và hệ thống ghi chú tự động.
  - Hỗ trợ nghiên cứu về AI trong xử lý giọng nói.

3.3.Google Colab

- **Là gì?**: Nền tảng lập trình trên đám mây cung cấp môi trường chạy Python miễn phí với GPU hỗ trợ AI/ML.
- **Ý nghĩa**: Giúp người dùng thực hành AI mà không cần đầu tư phần cứng mạnh.
- **Vai trò & Ứng dụng**:
  - Hỗ trợ chạy và huấn luyện mô hình AI trên nền tảng đám mây.
  - Được sử dụng rộng rãi để phát triển và thử nghiệm mô hình ngôn ngữ lớn (LLM).

3.4.Peft (Parameter Efficient Fine Tuning)

- **Là gì?**: Một kỹ thuật tinh chỉnh mô hình giúp giảm số lượng tham số cần tối ưu hóa khi fine-tuning LLM.

- **Ý nghĩa**: Giảm chi phí huấn luyện mô hình mà vẫn đạt được hiệu suất cao.
- **Vai trò & Ứng dụng**:
  - Giúp tinh chỉnh mô hình hiệu quả trên dữ liệu chuyên biệt mà không cần huấn luyện toàn bộ mô hình.
  - Ứng dụng trong điều chỉnh mô hình AI cho các lĩnh vực chuyên sâu như tài chính, y tế.

3.5.BitsAndBytes Library

- **Là gì?**: Thư viện hỗ trợ kỹ thuật lượng tử hóa giúp giảm kích thước mô hình AI.
- **Ý nghĩa**: Cho phép chạy mô hình lớn trên các thiết bị có tài nguyên hạn chế bằng cách giảm độ chính xác của trọng số mô hình.
- **Vai trò & Ứng dụng**:
  - Giúp triển khai mô hình AI trên điện thoại di động, IoT, và các hệ thống nhúng.
  - Hỗ trợ tối ưu hóa bộ nhớ khi triển khai mô hình lớn.

3.6.Gradio

- **Là gì?**: Công cụ giúp tạo giao diện người dùng dễ dàng cho AI mà không cần lập trình phức tạp.
- **Ý nghĩa**: Hỗ trợ kiểm tra, thử nghiệm mô hình AI nhanh chóng với giao diện trực quan.
- **Vai trò & Ứng dụng**:
  - Tạo giao diện thử nghiệm cho chatbot, công cụ nhận diện giọng nói, và mô hình xử lý ảnh.
  - Hỗ trợ nghiên cứu AI bằng cách cung cấp cách thức tương tác dễ dàng với mô hình.

3.7.AutoTokenizer & AutoModel

- **Là gì?**: Hai lớp trong thư viện Hugging Face Transformers giúp tải và sử dụng mô hình AI một cách tự động.
- **Ý nghĩa**: Giúp đơn giản hóa việc tích hợp mô hình vào ứng dụng AI.
- **Vai trò & Ứng dụng**:
  - Hỗ trợ load mô hình AI chỉ với một dòng lệnh mà không cần cấu hình thủ công.
  - Ứng dụng rộng rãi trong NLP, từ chatbot đến phân tích văn bản.

3.8.PyTorch & TensorFlow

- **Là gì?**: Hai framework học sâu phổ biến giúp huấn luyện và triển khai mô hình AI.
- **Ý nghĩa**: Cung cấp nền tảng mạnh mẽ cho việc phát triển các mô hình học sâu.

- **Vai trò & Ứng dụng**:
  - PyTorch thường được sử dụng cho nghiên cứu AI và mô hình thử nghiệm.
  - TensorFlow phổ biến trong sản xuất, hỗ trợ triển khai AI trên quy mô lớn.

3.9.Supervised Fine-Tuning (SFT)

- **Là gì?**: Kỹ thuật tinh chỉnh có giám sát giúp cải thiện hiệu suất của mô hình AI bằng cách huấn luyện trên dữ liệu gán nhãn.
- **Ý nghĩa**: Giúp mô hình AI hiểu và dự đoán tốt hơn nhờ dữ liệu có hướng dẫn.
- **Vai trò & Ứng dụng**:
  - Ứng dụng trong huấn luyện chatbot, hệ thống gợi ý, và AI cá nhân hóa.
  - Hỗ trợ tối ưu mô hình AI cho các ngành công nghiệp cụ thể như chăm sóc khách hàng.

3.10.Reinforcement Learning (RL)

- **Là gì?**: Phương pháp học máy trong đó AI học từ phản hồi của môi trường để cải thiện quyết định của mình.
- **Ý nghĩa**: Giúp AI thích nghi và đưa ra quyết định tối ưu hơn theo thời gian.
- **Vai trò & Ứng dụng**:
  - Ứng dụng trong huấn luyện trợ lý ảo, tối ưu hóa chatbot, và AI chơi game.
  - Hỗ trợ cải thiện các mô hình AI để phản hồi chính xác hơn dựa trên dữ liệu thực tế.