Multimodal Models with RAG

Course Overview

This 1-week intensive workshop is designed for developers, machine learning engineers, and researchers who are keen on exploring advanced applications combining the capabilities of retrieval-augmented generation (RAG) with vision-language models. Participants will learn how to utilize these technologies to enrich application functionalities with multimodal data understanding and generation, focusing on the integration of visual and textual information.

Learning Outcomes

Participants will:

- Deepen their understanding of vision-language models, including their architecture, functionality, and application areas.
- Learn how to effectively integrate vision-language models with RAG for enhanced multimodal data processing.
- Gain hands-on experience by implementing a project that utilizes these models to solve complex tasks involving both visual and textual inputs.

Recommended Prerequisites

- Strong proficiency in Python.
- Solid background in machine learning and deep learning.
- Familiarity with LLM and computer vision principles.
- Experience with deep learning libraries, preferably PyTorch or TensorFlow.

Detailed Curriculum Schedule

Week	Topics
Session 1: Vision-Language Models: Foundations and Applications (3 hrs)	• Introduction to Vision-Language Models (1 hr): Overview of the latest models (CLIP, DALL·E, etc.), their mechanisms, and use cases. • Integrating Vision-Language Models with RAG (1 hr): Exploring the concept of retrieval-augmented generation in the context of multimodal data. • Hands-on Exercise (1 hr): Setting up a simple vision-language model to perform tasks like image captioning or text-to-image retrieval.
Session 2: Project Implementation and Deep Dive into RAG (3 hrs)	• Advanced Techniques in Vision-Language Models (1 hr): Deep dive into fine-tuning and customizing models for specific applications. • Building a RAG-based Application (1 hr): Hands-on session on developing a multimodal application that uses RAG for generating responses based on both textual and visual inputs. • Mini Project and Wrap-up (1 hr): Participants will continue the project from session 1,, applying vision-language models and RAG to a chosen problem.

About FourthBrain

FourthBrain trains engineers, developers, data scientists, and leaders to make an impact in the Artificial Intelligence field, with our flexible, accessible education programs. We are training a new generation of engineers and leaders with more than just technical ability; they have an awareness and mindset of what is needed to succeed with AI. We are part of the AI Fund, founded by Andrew Ng.