Building with Multi-Modal Foundation Models

Course Overview

This workshop is a comprehensive 3-week program designed for machine learning engineers and team leads/managers. The workshop will explore the latest advancements in Large Multimodal Models (LMMs), with a focus on Foundation Models such as CLIP, LLaVa, Flamingo, Fuyu-8B, GPT-4V, Gemini, and others. Participants will gain theoretical knowledge and practical skills to develop applications that seamlessly integrate text, image, speech, and other modalities. The course structure includes a balance of theoretical discussions, hands-on coding sessions, and a culminating project to build a functional application showcasing participants' proficiency in building with Multi-Modal Foundation Models.

Learning Outcomes

By the end of the workshop, participants will be equipped with the ability to:

1. Understand the importance and applications of multimodal systems in various industries.
2. Master the fundamentals of multimodal training, using models like CLIP for contrastive language-image pre-training.
3. Apply multimodal models for tasks such as classification, text-based image retrieval, image generation, conversational assistant with multimodal inputs / outputs, etc.
4. Explore emerging research directions in the field of LMMs, including incorporating more data modalities, efficient training strategies, and generating multimodal outputs.

Recommended Prerequisites

- Proficiency in Python programming
- Solid understanding of machine learning fundamentals
- Familiarity with deep learning concepts and frameworks
- Intermediate knowledge of Natural Language Processing and Computer Vision

Detailed Curriculum Schedule

Week	Topics

Fundamentals of Multi-Modal Models and Applications
 	
(6 hrs)
	• Applications of multi-modal in various industries.
	• Exploring multimodal datasets and training objectives
	  ○ COCO-Captions, VQA, CMU-MOSEI, etc.
	• Hands-on session - Building a lightweight demo application with multi-modal capabilities (i.e. image captioning, visual question answering, etc.)
	  ○ Leveraging GPT-4V, Gemini, or open-source
	  ○ ImagineBind for search / retrieval use case
	• Deep dive into CLIP - understanding contrastive language-image pre-training
	• Introduction to Flamingo + Alternative Architectures
	• Hands-on session - OpenCLIP vs. OpenFlamingo - Comparative analysis applied to one specific use case
	
	Assignment #1:
	Design a concept for a multimodal application applicable to your company/domain/industry vertical and research available models or datasets that could be used to develop that capability.

Training & Fine-Tuning Multi-Modal Models (6 hrs)	• Building Multi-Modal datasets for fine-tuning • Leveraging pre-trained models to bootstrap data annotation • Hands-on session - Fine-tuning CLIP for a custom task • Experiment tracking **Assignment #2:** Evaluating performance with multi-modal models
Research Directions for LMMs & Best Practices (6 hrs)	• Incorporating more data modalities in LMMs • Multimodal systems for instruction-following • Adapters for more efficient multimodal training • Generating multimodal outputs - challenges and opportunities.

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.