Midterm and Final Projects

Full Stack Deep Learning, 2025

1. Multimodal AI for Smart Assistants

The rapid advancements in artificial intelligence have given rise to more intelligent and interactive virtual assistants. While existing AI assistants primarily rely on text or speech input, the next generation of AI assistants is shifting towards a multimodal approach—enabling machines to understand and process multiple data formats simultaneously.

Multimodal AI allows for a more natural, context-aware, and immersive human-computer interaction. Imagine an assistant that can not only answer spoken questions but also interpret images, generate real-time recommendations, and provide insightful responses based on different types of inputs. This capability is crucial for applications in education, healthcare, customer service, e-commerce, and accessibility for visually or hearing-impaired individuals. The assistant will be capable of performing various real-world tasks, such as:

- Answering questions based on images (e.g., "What is this object?" after uploading a picture).
- Responding to spoken commands while displaying relevant visual content.
- Summarizing and extracting insights from images, charts, or scanned documents.
- Providing real-time assistance for tasks like navigation, shopping recommendations, or troubleshooting based on multimedia inputs.

This project aims to bridge the gap between text, vision, and speech by developing a fully functional multimodal AI-powered smart assistant. The assistant will be designed to understand, analyze, and respond to a combination of text, speech, and image-based queries. Through this project, we will explore state-of-the-art multimodal architectures, integrate deep learning models for language, vision, and speech, and deploy a real-world application with a user-friendly interface.

By the end of the project, we will have a working prototype of a multimodal AI assistant that can perform a range of interactive tasks, setting the foundation for more advanced AI-driven personal assistants in the future.

Sample Code and Demo: LLaVA: https://github.com/LLaVA-VL/LLaVA-NeXT

2. AI for Healthcare Diagnostics

Medical imaging plays a critical role in healthcare, helping doctors detect diseases early and make informed treatment decisions. However, analyzing medical images requires extensive expertise, and misdiagnosis can have severe consequences. AI-powered diagnostics have shown promise, but traditional deep learning models often lack contextual understanding, making them difficult to integrate into real-world clinical workflows.

Large Language-Vision Models (LLVMs) offer a transformative approach to medical diagnostics by combining computer vision and natural language processing (NLP). Unlike standalone image-based AI models, LLVMs can:

- Analyze medical images in conjunction with clinical notes and patient history.
- Answer complex medical queries related to an image or report.
- Generate structured radiology reports with explanations.
- Provide human-like reasoning and justifications for AI-driven diagnoses.
- This project aims to develop a multimodal AI diagnostic assistant that enhances medical image analysis, clinical decision-making, and report generation. The model will be trained on large medical datasets (e.g., MIMIC-CXR, NIH ChestX-ray14, CheXpert) and fine-tuned for healthcare applications.

This project bridges the gap between AI-driven medical imaging and real-world clinical decision-making, making healthcare diagnostics more accurate, accessible, and trustworthy.

Sample Code and Demo: LLaVA-Med https://github.com/microsoft/LLaVA-Med

3. Image-to-Video Generation

The ability to generate realistic videos from static images represents a major breakthrough in generative AI. Traditional video creation requires manual animation, rendering, or filming, making it time-consuming and resource-intensive. However, image-to-video generation powered by deep learning can automate this process by predicting motion and generating temporally coherent video sequences from a single image input.

Recent advancements in diffusion models, generative adversarial networks (GANs), and vision transformers have enabled AI models to synthesize high-quality motion, interpolate frames, and animate objects realistically. These models can generate consistent, dynamic motion patterns that match real-world physics, whether facial expressions, natural elements like water and fire, or complex human actions.

This project aims to build an AI-driven image-to-video generator that allows users to upload static images and transform them into realistic, high-resolution videos. The system will leverage a multimodal approach, integrating text prompts, video diffusion models, and neural motion prediction networks to enable customizable, user-guided video generation.

By the end of this project, the goal is to develop a functional, scalable, and interactive image-to-video generation tool that can be used in various applications, from digital content creation to AI-driven animations and scientific visualizations.

Sample Code and Demo:

- Video Stable Diffusion: https://huggingface.co/spaces/multimodalart/stable-video-diffusion
- Diffusion Forcing Transformer: https://huggingface.co/spaces/kiwhansong/diffusion-forcing-transformer

4. Text-to-Image Generation

Text-to-image generation is one of the most revolutionary advancements in Generative AI, allowing users to transform textual descriptions into stunning, high-resolution images with minimal effort. Traditional image creation requires manual design, drawing, or photography, making it time-consuming and skill-intensive. However, deep learning-powered text-to-image models have made it possible to generate complex, visually appealing content within seconds.

Advances in diffusion models, transformer-based vision architectures, and contrastive learning (CLIP) have significantly improved the semantic accuracy, realism, and stylistic diversity of AI-generated images. These models understand natural language prompts and translate them into meaningful visual representations while maintaining fine details, artistic control, and photorealism.

This project aims to develop a Text-to-Image Generation system that allows users to input detailed prompts and receive AI-generated images tailored to their requests. The system will integrate state-of-the-art deep learning techniques, ensuring high fidelity, control over style and composition, and scalability for real-world applications.

By the end of this project, we will have a fully functional, AI-powered text-to-image generation platform, capable of creating high-quality, customizable images for various use cases, from creative content generation to digital art and scientific visualizations.

Sample Code and Demo: Stable Diffusion https://huggingface.co/spaces/stabilityai/stable-diffusion-3.5-large

5. Virtual Try-on for Clothing Fashion

The rise of e-commerce and digital fashion has transformed how people shop for clothes. However, one of the biggest challenges in online shopping is the inability to try on clothing before purchasing, leading to uncertainty, high return rates, and customer dissatisfaction.

AI-powered Virtual Try-On (VTO) provides a game-changing solution, enabling users to see themselves wearing different outfits digitally. Using Generative AI and deep learning, this system can create a realistic virtual dressing room, where users can upload their images, select clothing items, and receive a high-fidelity visualization of how they would look in different outfits.

Recent advancements in deep learning-based image synthesis, neural rendering, and computer vision allow AI models to adapt clothing to various body shapes, simulate realistic fabric textures, and generate high-quality try-on experiences. Leading fashion and e-commerce companies are investing in AI-powered virtual try-on systems to revolutionize the shopping experience and reduce return rates, making the industry more sustainable.

This project aims to develop a cutting-edge Virtual Try-On system that combines computer vision, deep learning, and generative AI to offer a seamless and interactive fashion experience. By the end of the project, we will have a functional AI-powered virtual dressing room, capable of real-time clothing simulation, accurate outfit visualization, and integration with online shopping platforms.

Sample Code and Demo: https://huggingface.co/spaces/HumanAIGC/OutfitAnyone

6. AI-Powered Smart OCR Scanner and Translation Deployed on the Edge

OCR (Optical Character Recognition) is a crucial technology for digitizing text from physical documents, signs, and handwritten notes. However, most OCR and translation solutions rely on cloud-based APIs, making them dependent on internet access, slow response, and privacy-sensitive.

This project aims to build an on-device AI-powered OCR and translation system that eliminates cloud dependencies while delivering fast, accurate, and secure results. By leveraging edge AI models, the system will be capable of real-time text detection, recognition, and translation on mobile devices and embedded AI platforms.

With the advancement of lightweight deep learning models like Transformer-based OCR (TrOCR) and optimized neural machine translation, we can now deploy high-performance OCR and translation systems on low-power devices. This allows users to extract and translate text from documents, street signs, business cards, restaurant menus, or handwritten notes—all offline and in real-time.

Sample Code and Demo:

- OCR Detection: https://github.com/microsoft/unilm/tree/master/trocr
- Translation: https://huggingface.co/docs/transformers/en/tasks/translation