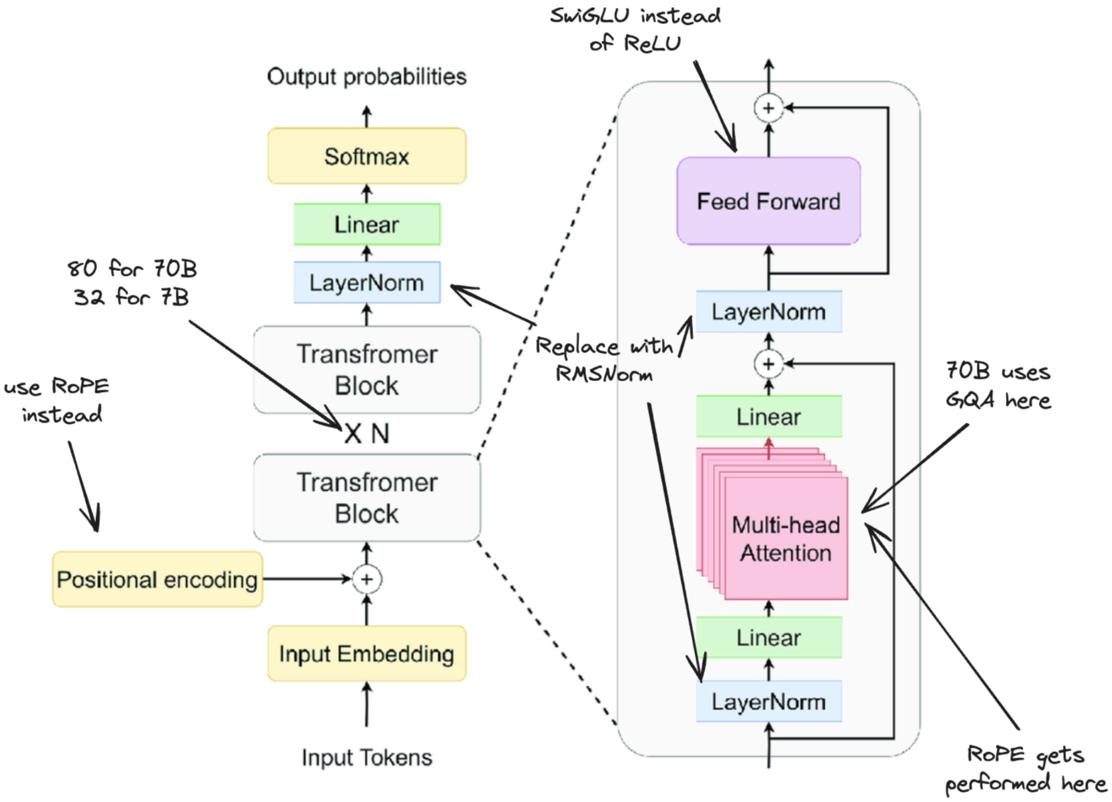
Image Understanding in LLMs

Project Description. The impressive success of large language models (LLMs) has sparked an increased need to manage multiple modalities beyond language. As a result, Large Multimodal Models (LMMs) such as GPT-4v, Gemini, DeepSeek-VL, and Qwen2-VL have emerged. These models can understand and act on instructions involving *both vision and language*, i.e., they enable users to upload an image and talk with the LLM about it.

In principle, multimodal Transformers, e.g., CLIP and BLIP, are designed to handle both text and image input. These models process both visual and textual data in a joint space. This allows them to understand the text and connect it to visual representations. The general framework is as follows: *i) image features* are first extracted through a vision transformer such as ViT which converts visual data into embeddings, *ii) textual input* is processed by a language model, which converts the text into its own embedding, and then *iii) both embeddings* are processed together by a shared transformer architecture or through cross-attention mechanism. However there are some architectural details that differentiate these models from each other.



In this project, we are interested in demystifying the architecture of *DeepSeek-VL* and *Qwen2-VL*. In particular, we will explore the *Mixture-of-Experts (MoE) framework* combined with the *Multi-Head Latent Attention* mechanism. This advanced transformer-based architecture enables the vision-language model to handle complex tasks with high accuracy and speed while being cost-effective and achieving state-of-the-art results.

Project Type. 60% Engineering, 40% Research.

Internship Batch. Batch 1 from May 11 to July 10, 2025.

Duties/Activities Include: Read and analyze research papers in AI and machine learning. Examine and update source codes of advanced machine learning models, focusing on LLMs and multimodal LLMs. Train, fine-tune, and deploy machine learning models on large-scale clusters as well as on common platforms such as Huggingface. Implement new ideas to improve the performance and accuracy of machine learning models. Write code to collect and process various image, video, and text datasets. Analyze the performance of machine learning models using multiple benchmarks. Write reports and presentations explaining the source code, datasets, and results.

Required Skills. Strong programming skills, especially in Python. Ability to work with large code bases. Strong mathematical background at the bachelor's or master's level, including good knowledge of advanced topics in Calculus and Linear Algebra.

Preferred Intern Academic Level. B.Sc. (3rd year or 4th year) or MS student.

Learning Opportunities. Students will be exposed to exciting research in Large Language Models (LLMs). They will examine state-of-the-art models and learn how such models work. Students will learn by implementing, analyzing, and updating the architectures of recent models such as DeepSeek-VL and Qwen2-VL. Students will learn how to handle datasets in various modalities, including images, videos, and texts. Students will gain experience in training and deploying LLMs on large-scale compute clusters with powerful GPUs.

Expected Team Size. It is preferable to have a team of 2 interns.

Mentors. Dr. Abdelkader Baggag, Dr. Mohamed Hefeeda