Project title

Mixture of Experts for Vision-Language Models (VLMs)

Supervision team

Main Supervisor: Varun Ojha < Varun.Ojha@newcastle.ac.uk >
Co-supervisors: Huizhi Liang <Huizhi.Liang@newcastle.ac.uk >

Research project

The integration of vision and language presents a significant challenge in artificial intelligence (AI), leading to the development of Vision-Language Models (VLMs). These models facilitate various tasks, including image captioning, visual question answering (VQA), cross-modal retrieval, and embodied reasoning. Large-scale models such as CLIP, Flamingo, and GPT-4V have showcased impressive multimodal reasoning capabilities; however, these advancements come with substantial computational demands and limited interpretability. A promising approach to overcome these limitations is the Mixture of Experts (MoE) framework. This framework involves the selective activation of different expert subnetworks based on the input, providing both scalability—by allowing sparse activation of large networks—and specialization—by permitting experts to concentrate on specific modalities, domains, or reasoning strategies. While MoEs have been extensively studied in natural language processing (NLP), their application to vision-language integration has not been thoroughly explored. This PhD project aims to investigate how MoE architectures can be utilized to enhance the efficiency, adaptability, and interpretability of Vision-Language Models.

Applicant skills/background

This project requires skills in programming in python and machine learning research skills. The project welcomes researchers from disciplines of mathematics, engineering, physics, computer science, electronics.

References

• Radford et al., 2021. *Learning Transferable Visual Models From Natural Language Supervision (CLIP).*

• Alayrac et al., 2022. *Flamingo: A Visual Language Model for Few-Shot Learning.*