EE4052 Task 3: Project Abstract

Student Name:	**Jaikrishna Reddy Gundala**	Id Number:	**24131989**

Project Title: **Layer-wise Analysis of Cross-Modal Attention in Vision Language Models for VQA**

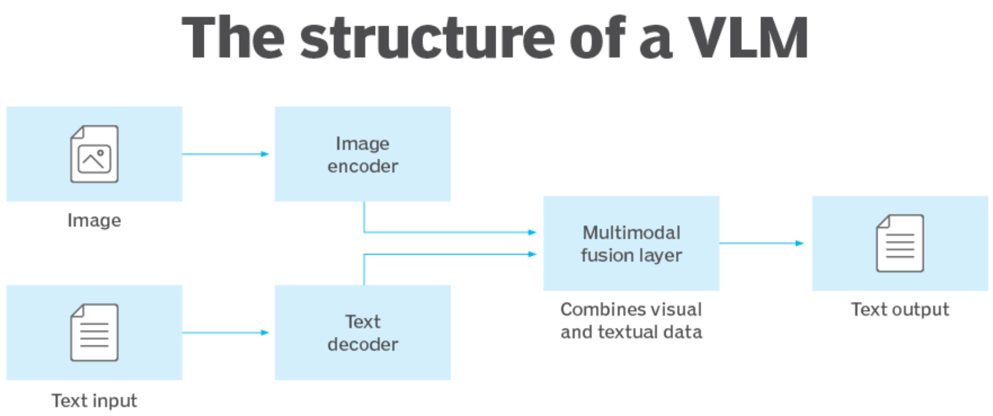
Project Areas: Computer vision, Natural Language Processing, Vision Language Models Visual Question and Answers, Transformers & Attention, Deep Learning, AI.

Description:

Transformer-based Vision-Language Models (VLMs) process both visual and textual inputs using stacked transformer encoders, enabling tasks like Visual Question Answering (VQA). However, how these layers distribute attention across modalities remains unclear. This project investigates layer-wise attention patterns in VLMs to understand how different layers specialize in vision-language interactions.

Existing research has explored attention in transformers, but most studies focus on overall patterns rather than layer-wise differences in Vision-Language Models (VLMs). While attention variations across layers are well-documented in NLP and vision models, their specific role in multimodal tasks like VQA remains underexplored.

This project analyses layer-wise attention in Vision-Language Models (VLMs**)** for Visual Question Answering (VQA). It aims to uncover how different transformer layers process visual vs. textual information and how attention patterns vary across question types **("what," "where," "why")**. The findings will enhance model interpretability and help optimize VLM architectures for improved performance.



Resources required:

Datasets and Model:

VQA v2, GQA, CLIP, ViLT, VisualBERT.

Hardware & Computational Resources:

NVIDIA A100/V100 GPUs, 64 GB RAM, external SSDs, network storage, High-Performance Computing (HPC).

Software & Development Tools:

PyTorch, Hugging Face Transformers, Matplotlib, Seaborn, NumPy, OpenCV.

Reference:

1. *Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, **"Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering,"** Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.https://arxiv.org/abs/1612.00837*
2. *T. Wolf et al., **"Transformers: State-of-the-Art Natural Language Processing,"** Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2020.https://arxiv.org/abs/1910.03771*
3. *A. Vaswani et al., **"Attention Is All You Need,"** Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017. https://arxiv.org/abs/1706.03762*