



***Layer-wise Analysis of Cross-Modal Attention in Vision
Language Models for Visual Question and Answer***

LK489 - Master of Engineering in Computer Vision and
Artificial Intelligence

Interim Project Report

Student Name: **JAIKRISHNA REDDY GUNDALA**

Student ID: **24131989**

Supervisor Name: **Tony Scalan**

Date: **03/05/2025**

Abstract:

This project investigates how Vision-Language Models (VLMs), Particularly ViLT, manage attention mechanism across their layers while handling both images and text in Visual Question Answering (VQA) tasks. However these models are good at processing multiple types of data at once, it's remained unclear that how each layer contributes to understand visual and textual information. By analysing attention patterns across multiple layers and observing how they shift attention depending on the type of questions like ("what," "where," or "why") hope this project aims to understand better on how these models think. The goal is to make these systems more interpretable and help improve their design for better performance. All the experiments will use the VQAv2 dataset and be run using ViLT, from Hugging face transformers on Google Colab's GPU.

Keywords: Vision-Language Models, Visual Question Answering, Transformers, Attention Analysis, Deep Learning, ViLT, Computer Vision, NLP.

Declaration

This interim report is presented in part fulfilment of the requirements for **the LK489 Master of Engineering in Computer Vision and Artificial Intelligence** Masters Project.

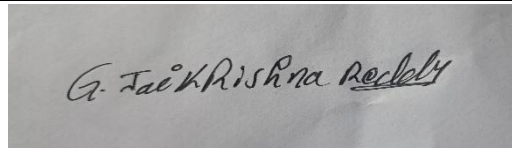
It is entirely my own work and has not been submitted to any other University or Higher Education Institution or for any other academic award within the University of Limerick.

Where there has been made use of work of other people it has been fully acknowledged and referenced.

Name

G JAIKRISHNA REDDY

Signature



Date

03/05/2025

Introduction

Project Aims and Objectives

The main objective of my project is to investigate how attention is distributed across different layers of Vision-Language Models during Visual Question Answering tasks. The research seeks into:

- Analyse layer wise specific attention behaviours in ViLT.
- Understand how attention varies based on question types.
- Enhance the interpretability of VLMs.
- Offer insights that may optimize future multimodal model architectures.

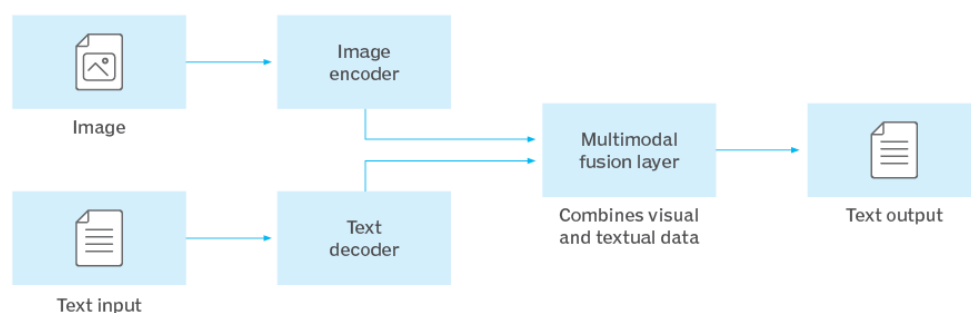
Motivation and Problem Statement

Modern Vision-Language Models, such as ViLT, excel in tasks that combine vision and language. However, their internal works, particularly how visual and textual inputs interact through stacked transformer layers, remain largely opaque. Understanding layer-wise attention specialization could reveal inefficiencies or overlooked patterns, leading to more efficient and interpretable models.

Societal/Technological Impact

Interpretability in AI, especially in multimodal models, is a rising concern across industries like healthcare, autonomous systems, and assistive technologies. VQA models that clearly explain their decision-making could boost trust and reliability, particularly in high-stakes fields like medical diagnosis from images and captions (Lu et al., 2019). Further, optimizing VLM architectures has technological benefits, reducing computational costs and enabling deployment on resource-constrained devices.

The structure of a VLM



Literature Survey

1. Vaswani et al., 2017 – “Attention is All You Need”

In their groundbreaking paper **Attention is All You Need**, Vaswani et al from google. introduced the Transformer architecture, which is now become the foundation for most of the modern NLP and vision-language models. Instead of relying on recurrence or convolution, the Transformer uses a self-attention mechanism to understand the relationships between input tokens—like words in a sentence or patches in an image. This means every token can directly consider every other token, regardless of position, making the model highly flexible and efficient. The architecture is built from layers of multi-head self-attention and feedforward networks. Each attention head can learn to capture different aspects of the data, such as meaning, grammar, or position, resulting in richer and more powerful representations.

The Transformer framework introduced by Vaswani et al. directly influences models like ViLT and other vision-language systems. In ViLT, images are divided into patches and tokenized, while text is converted into embeddings. Both are then fed into the same Transformer layers, allowing the model to process and relate visual and textual information without needing separate encoders for each. Self-attention plays a key role here, enabling the model to learn cross-modal relationships. For this project, this paper provides both the conceptual and architectural foundation. this focus on analysing attention patterns in each Transformer layer—and how different heads behave in multimodal tasks (e.g., attending to image patches in "what" questions and text in "why" questions)—builds directly on the mechanisms introduced in this work. Additionally, the interpretability offered by attention maps supports your investigation into how attention specialization develops across layers in ViLT.

2. Agrawal et al., 2015 – “VQA: Visual Question Answering”

Agrawal et al.’s paper introduced the Visual Question Answering (VQA) task along with the first large-scale dataset, VQAv1, marking a key moment in the evolution of multimodal AI. Their approach combined Convolutional Neural Networks (like VGGNet or ResNet) to extract features from images and Recurrent Neural Networks—specifically LSTMs—to encode the question. These visual and textual features were then merged and passed through a classifier to generate an answer. The paper highlighted the open-ended nature of VQA, where answers range from simple yes/no responses to more complex ones involving objects, colours, or reasoning.

Although the model didn’t use attention mechanisms or Transformers, it set the stage for framing VQA as a multimodal reasoning challenge—requiring models to understand and integrate both visual and textual inputs. It also brought attention to key issues like language biases and dataset imbalances, which later led to improvements in dataset design, such as VQAv2.

For your project, this work serves as the foundation for the VQA task itself, providing both the benchmark dataset and evaluation metrics. The insight that different question types rely on different kinds of visual grounding—for example, “where” questions needing spatial understanding—directly supports your plan to analyse attention behaviour in ViLT by question type. Exploring how ViLT balances attention between text and image based on question categories builds directly on this foundational work.

3. Wang et al., 2020 – “Visual Question Answering: Attention Mechanism, Datasets, and Future Challenges”

This review paper offers a comprehensive overview of how attention mechanisms have evolved in Visual Question Answering (VQA) models. It traces the shift from early methods like simple pooling to more advanced cross-modal attention systems. The authors classify attention into two main types: soft attention, where the model assigns probabilities to different image regions, and hard attention, which selects specific regions in a non-differentiable way. They also cover co-attention mechanisms, where the model learns to attend to both image and text simultaneously. In this paper many important models are discussed, including Stacked Attention Networks (SANs), Hierarchical Co-Attention Networks, and more recent Transformer-based architectures like LXMERT and ViLBERT.

The main theme of the paper is to analyse how attention mechanisms help bridge the gap between visual and textual information, allowing models to focus on parts of an image that are relevant to a given question. The authors also stress that attention is not uniform different layers and heads often specialize in different tasks, such as recognizing objects or reasoning about relationships. The review highlights the ongoing challenge of making attention-based models interpretable, especially in sensitive domains like healthcare or law.

Your project builds directly on these ideas. By analysing attention in ViLT at the layer and head level and breaking it down by question type you’re addressing a gap identified in the paper: the need to understand how attention works inside models, not just whether it works. Your use of visualizations and stratified analysis is also closely aligned with the review’s recommendation to develop more transparent and interpretable VQA systems.

4. Kodali and Berleant, 2022 – “Recent, Rapid Advancement in Visual Question Answering: A Review”

Kodali and Berleant’s paper offer a contemporary overview of the VQA research landscape, with a focus on the transition to transformer-based architectures. It categorizes VQA models into three main types: modular, where components are explicitly structured for parsing and reasoning (e.g., MAC networks); monolithic, where a single end-to-end model handles everything (e.g., ViLT); and hybrid, which combine structured and holistic approaches. The review also examines recent datasets (e.g., GQA, TDIUC, OK-VQA) and how they test different reasoning abilities, from factual recall to common-sense and causal inference.

The paper particularly highlights how models like ViLT, UNITER, and LXMERT leverage transformers to process multimodal inputs with minimal pre-processing. However, it also critiques the black-box nature of these models, noting that attention weights and internal dynamics are often poorly understood. The authors point to layer-wise behaviour and attention specialization as promising areas for future research—exactly the focus of your project.

This review validates the importance of your work by situating it in the broader trend toward interpretable, efficient, and deployable VQA models. By investigating how ViLT distributes attention across layers, and how different question types affect attention allocation, you provide a detailed probe into a problem this paper identifies but does not solve. Your use of visualization tools and stratified analysis addresses the call for more transparent VLM behaviour.

Conclusion of Literature Survey

My literature review highlights few substantial advancements in Vision Language Models (VLMs) and their applications for Visual Question Answering (VQA). A key turning point was the introduction of the Transformer based architecture introduced by Vaswani et al. (2017), which enabled more effective joint processing of visual and textual inputs and laid the groundwork for many modern multi-modal models. The VQA task itself was first formalized by Agrawal et al. (2015), establishing a benchmark for evaluating a model's ability to reason across modalities. Later reviews, such as those by Wang et al. (2020) and Kodali & Berleant (2022), have traced the evolution of attention mechanisms and emphasized the growing need for both interpretability and computational efficiency.

A common thread across these works is the central role of attention mechanisms in aligning vision and language. While earlier models relied on dedicated cross-modal attention modules, newer architectures like ViLT integrate both modalities within a shared Transformer backbone. This shift simplifies the architecture but also makes it even more important to understand how attention is distributed between visual and textual inputs across layers.

Despite the availability of many strong models and big scale datasets, still we can find a noticeable gap in our understanding of how attention behaves inside these systems. Few studies have taken a layer-by-layer look at attention in unified architectures like ViLT, especially in the context of different question types (such as “what,” “where,” or “why”). There is also limited work on visualizing and interpreting why a model chooses to attend to certain regions or words during VQA tasks.

Identified Gaps:

- Limited analysis of layer-wise attention behaviour in vision-language models like ViLT.
- Lack of insight into how attention patterns vary with different types of questions.

- Minimal exploration of how cross-modal specialization emerges in shared Transformer layers.
- A shortage of visual tools to help interpret attention and internal decision-making in VLMs.

Future Scope and Relevance:

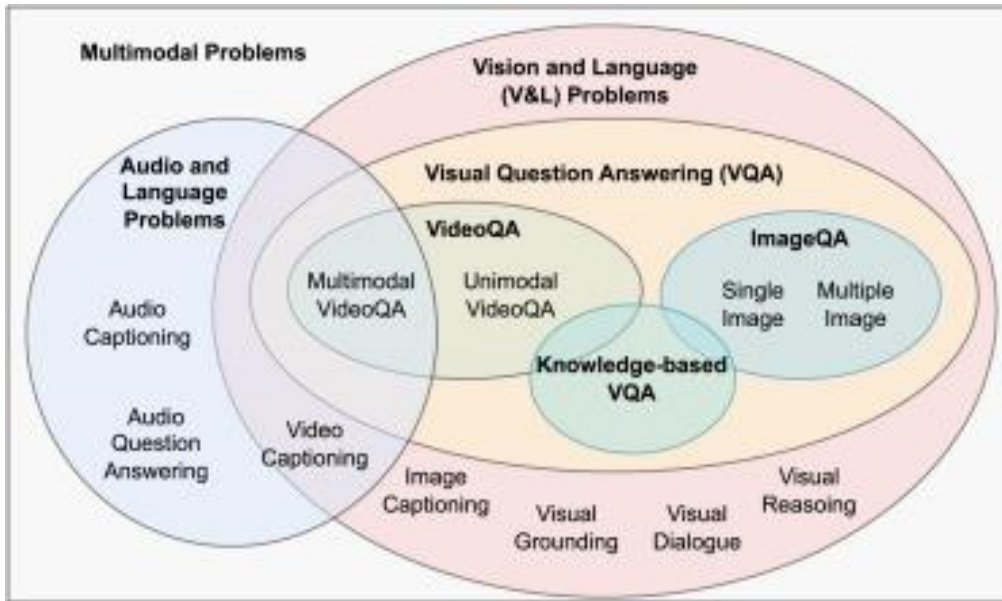
This project seeks to address these gaps through a detailed analysis of attention maps across all layers of ViLT, with a focus on how attention shifts based on question type. It will also generate visualizations such as heatmaps and attention graphs—to support interpretability. The goal is to contribute to a deeper understanding of how VLMs reason across modalities and to support the development of models that are not only accurate but also transparent and easier to trust.

Beyond interpretability, these insights could also inform practical improvements, such as identifying which attention heads or layers are most critical for specific tasks. This might help in model pruning, efficiency gains, and more lightweight architectures. In doing so, the project aligns with broader efforts toward building responsible, explainable, and resource-efficient AI systems especially important in high-stakes applications like healthcare, autonomous systems, and assistive technology.

Methodology

The ViLT model uses a unified transformer encoder where both text tokens and image patches are fed simultaneously. Unlike traditional VLMs like LXMERT, which use separate backbones for vision and text, ViLT allows easier tracking of cross-modal attention.

- **Attention Map Extraction:** Intercept attention outputs at each layer during inference.
- **Analysis:** Quantify modality-specific attention weights (textual vs visual) per layer.
- **Categorization:** Segment results by question types ("what", "where", "why").
- **Visualization:** Create heatmaps and attention distribution graphs.



Outline of Designs

Tools and Setup

To carry out this project effectively, I've chosen tools that are well-established in the machine learning community:

Python is the main programming language due to its ease of use and strong support for deep learning research.

PyTorch is used for building and working with neural networks, offering flexibility and great support for custom model inspection—especially useful for attention extraction.

Hugging Face Transformers provides pre-trained models like ViLT, making it straightforward to fine-tune and analyse them for VQA tasks.

Matplotlib and Seaborn are used for visualizing the attention patterns across layers, helping to interpret what the model is focusing on during different types of questions.

Google Colab Pro with GPU gives access to high-performance computing resources without needing local hardware, allowing me to train and run attention analyses efficiently.

Dataset

I'll be using a **subset of the VQAv2 dataset**, which consists of images paired along with the open-ended questions' dataset is a standard benchmark in the VQA community and is ideal for analysing how models handle different types of questions like "what," "where," and "why." Using a curated subset keeps the experiments manageable while still covering enough variety to provide meaningful insights.

Model

The core model for this project is **ViLT (Vision-and-Language Transformer)**. It's designed to handle both images and text in a unified transformer architecture, which makes it easier to analyse how cross-modal attention is distributed. Since it doesn't rely on heavy visual feature extractors like CNNs, ViLT is more efficient and better suited for tracking attention across layers. I'm fine-tuning it on the VQA subset to adapt it specifically to the task at hand.

Action Plan

Gantt Chart Action Plan (May 18 – August 25)

Task No.	Task Description	Start Date	End Date	Duration (weeks)
1	Model Setup and Data Preprocessing	May 18	May 31	2 weeks
2	Initial Training	June 1	June 14	2 weeks
3	Layer-Wise Attention Analysis	June 15	June 28	2 weeks
4	Attention Analysis for Question Types	June 29	July 12	2 weeks
5	Refining Analysis and Layer Roles	July 13	July 26	2 weeks
6	Optimizing Model and Refinement	July 27	Aug 9	2 weeks
7	Final Report and Presentation	Aug 10	Aug 25	2 weeks

Project Requirement Elicitation

Hardware Requirements

- Platform: Google Colab Pro
- GPU: Tesla T4 or better

Compute Requirements

- Moderate workload: a single GPU is sufficient
- Occasional memory upgrades may be needed for larger dataset subsets or visualizations

Software Requirements

- **Core Libraries:** PyTorch (latest stable), Hugging Face Transformers
- **Additional Tools:** matplotlib, seaborn, numpy, pandas

Data Requirements

- Dataset: VQAv2 (downloaded from the official site)
- Preprocessing: Data preformatted for ViLT (image patches + text embeddings)

Reference

- Vaswani et al., "Attention is all you need," arXiv:1706.03762, Jun. 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>.
- Agrawal et al., "VQA: Visual question answering," arXiv:1505.00468, May 2015. [Online]. Available: <https://arxiv.org/abs/1505.00468>.
- M. Tang, R. Wang, S. Lu, A. Alsanad, and L. Zhang, "Visual Question Answering: Attention Mechanism, Datasets, and Future Challenges," in *Proc. 9th Int. Conf. Cyber Security Inf. Eng. (ICCSIE)*, Kuala Lumpur, Malaysia, Dec. 2024, pp. 1–8. doi: [10.1145/3689236.3691498](https://doi.org/10.1145/3689236.3691498).
- V. Kodali and D. Berleant, "Recent, rapid advancement in visual question answering architecture: A review," arXiv:2203.01322, Mar. 2022. [Online]. Available: <https://arxiv.org/abs/2203.01322>.