

Towards a performance analysis on pre-trained Visual Question Answering models for autonomous driving

Kaavya Rekanar*, Ciarán Eising*, Ganesh Sistu†, Martin Hayes*

*University of Limerick, †Valeo Vision Systems

*firstname.lastname@ul.ie, †firstname.lastname@valeo.com

Introduction

Visual Question Answering (VQA) is the process of generating natural language responses to open-ended questions by leveraging visual information derived from an image.

Why VQA? Contextual understanding, enhancing human-machine interaction, facilitating adaptive decision-making, and contributing to safety and error handling.

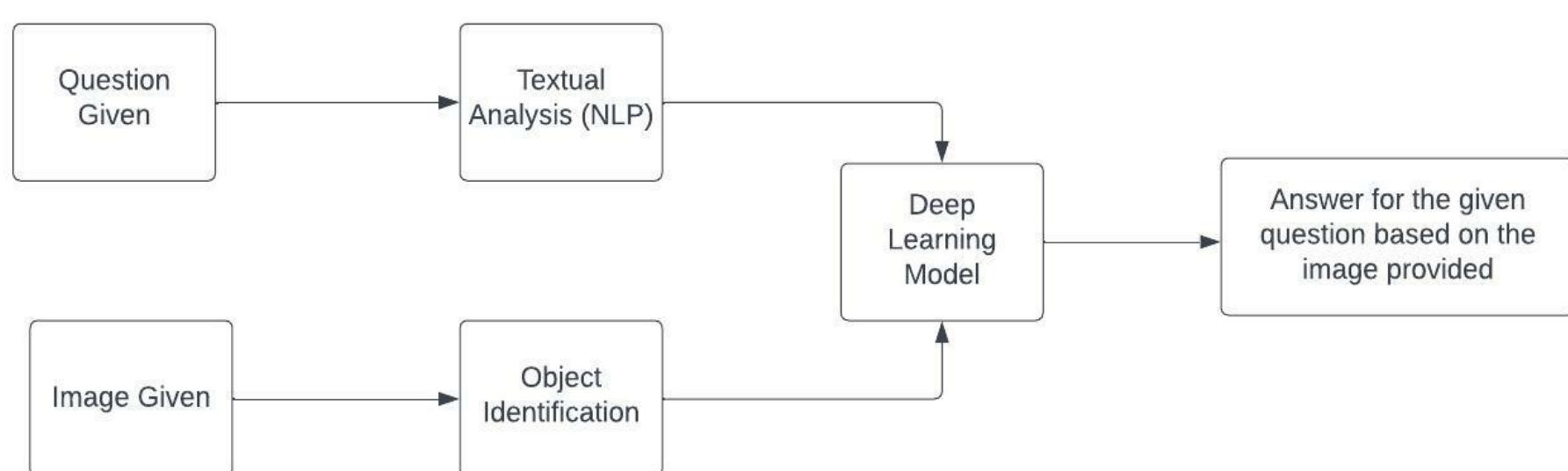


Fig 1: Architecture of VQA models

Related Work

VQA models incorporate multimodal architectures that utilize transformers to handle the fusion of visual and textual modalities.

Aspect	Processing Mechanism	Utilization in Multimodal Models
Early Fusion	Simultaneous processing of visual and textual inputs	Enables joint representation learning, capturing relationships between modalities from the start
Cross-Modal Attention	Information exchange and alignment between modalities	Enhances multimodal understanding and integration, capturing correlations and interactions
Late Fusion	Separate processing of visual and textual inputs	Facilitates individual representation learning, followed by fusion to capture interdependencies

Table 1: Utilization of Transformers in Multimodal Models

Methodology

- Overview of the analysis conducted on three models: ViLBERT [1], ViLT [2], LXMERT [3].
- Focused specifically on their performance in the domain of Visual Question Answering (VQA) with a strict focus on driving scenarios.

Model	Approach	Description
Vision and Language BERT (ViLBERT)	Early Fusion	ViLBERT extends BERT with a co-attention mechanism, integrating vision-attended language features into visual representations. It enables joint reasoning about text and images for visual grounding.
Vision-and-Language Transformer (ViLT)	Cross Modal Attention	ViLT aligns visual and textual features using cross-modal attention mechanisms. It generates joint representations through a visual encoder and a language encoder.
Vision-Language Transformer With Weakly-Supervised Local-Feature Alignment (VoLTA)	Late Fusion	VoLTA combines visual and textual information with a weakly supervised local-feature alignment mechanism. It focuses on relevant image parts without precise object-level annotations.

Table 2: Overview of Multimodal Models experimented

Results







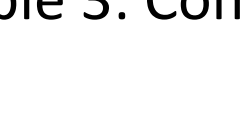
Category	Image	Questions	CV Experts	ViLBERT	ViLT	LXMERT
Dark setting		what are the contents of the image?	An Accident	Trucks	Cars	Cars
		what should the driver do?	Slow down and go left	Run	Stop	Go
Light Setting		what are the contents of the image?	Traffic Lights	Clouds	Traffic Lights	Power Lines
		what should the driver do?	Stop	Sleep	Stop	Stop
Parking		what are the contents of the image?	Parking Lot	Clouds	Cars	Cars
		Can the driver park here?	Yes	No	No	Yes
Signboard		what are the contents of the image?	Road	Paint	Traffic Lights	Cars
		what should the driver do?	Stop	Sleep	Stop	Go
Pedestrian Crossing		what are the contents of the image?	People crossing the road	Clouds	Buildings	People
		what should the driver do?	Stop	Run	Stop	Stop
Traffic		what are the contents of the image?	Traffic	Trucks	Buses	Buses
		what should the driver do?	Go	Run	Stop	Go
Accident		what are the contents of the image?	An Accident	Clouds	Cars	Cars
		what should the driver do?	Stop	Stop	Stop	Stop

Table 3: Comparison of Responses: Computer Vision Experts vs. Selected Models

Summary

- ViLBERT demonstrates a lack of comprehension regarding the question "what are the contents of the image?", as it consistently provided the answer "nothing".
- ViLT exhibits a certain level of capability in generating answers based on the provided images. Upon investigation, it becomes apparent that the model comprehends the question to some extent, and its object identification performance surpasses that of ViLBERT.
- LXMERT demonstrates commendable performance in answering questions within a driving context. Although the model exhibits excellent object identification capabilities, the accuracy of its answers requires refinement.
- The ability to accurately comprehend and respond to user queries in real-time scenarios remains a crucial aspect of enhancing the interaction between drivers and vehicles.

References

- [1] Lu, J., Batra, D., Parikh, D. and Lee, S., 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- [2] Kim, W., Son, B. and Kim, I., 2021, July. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning* (pp. 5583-5594). PMLR.
- [3] Tan, H. and Bansal, M., 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Preprint and Supplementary material



Acknowledgements

This work has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI 18/CRT/6049

HOST INSTITUTIONS



FUNDED BY

