Visual Question Answering with Scene Graphs

Saurabh Madake
M. Tech. (Computer Engineering) / (Data Science)
COEP Technological University (COEP Tech)
Pune, India - 411005
madakesb23.comp@coeptech.ac.in

Archana Patil
Assistant Professor, Department of Computer & IT
COEP Technological University (COEP Tech)
Pune, India - 411005
abp.comp@coeptech.ac.in

***Abstract* – This research focuses on developing a Visual Question Answering (VQA) system leveraging transformer architectures, particularly LXMERT (Learning Cross-Modality Encoder Representations from Transformers), to bridge the gap between visual and textual data. VQA is an AI-complete task that requires answering questions based on the content of an input image, presenting significant challenges in computer vision and natural language processing. Our work employs the VQA v2.0 dataset and Visual Genome for training and evaluation, integrating question-answer pairs with scene graphs, relationships, and attribute data. This study advances the field by pushing the boundaries of multi-modal deep learning and exploring the potential of transformers in addressing complex AI-complete tasks.**

Key Words: Transformers, VQA, LXMERT, Visual Genome

I. INTRODUCTION

Visual Question Answering (VQA) represents a cutting-edge challenge at the intersection of computer vision and natural language processing, aiming to bridge the gap between visual and textual understanding. As an AI-complete task, VQA demands sophisticated reasoning capabilities, making it a significant problem in artificial intelligence research. The goal is to develop systems capable of interpreting visual content and responding accurately to questions about it, with potential applications ranging from assistive technologies for visually impaired individuals to enhancing human-computer interaction. By enabling machines to answer questions based on visual inputs, VQA pushes the boundaries of machine comprehension and has profound implications for real-world AI applications.

Our research focuses on leveraging transformer-based models, particularly LXMERT, to address the complexities of VQA. Traditional approaches relied on separate processing pipelines for visual and textual data, often leading to integration challenges. However, transformer architectures such as LXMERT process multimodal data simultaneously, significantly improving VQA accuracy and performance. By leveraging attention mechanisms, LXMERT enables deep alignment between visual content and language, making it highly effective for this task. Implementing LXMERT within robust frameworks such as PyTorch and Hugging Face allows for an in-depth exploration of the model’s capabilities, limitations, and potential improvements in VQA.

To support our analysis, we utilize well-established datasets such as VQA v2.0 and Visual Genome, which provide diverse images, questions, and associated metadata essential for model training and evaluation. These datasets offer varied scenes, question types, and visual relationships, contributing to a more comprehensive assessment of VQA performance. Through this study, we aim to evaluate LXMERT’s effectiveness in handling VQA tasks while providing insights into the evolution, current challenges, and future directions of VQA technology. By refining transformer-based methodologies, our research contributes to advancing multimodal AI, enhancing model accuracy, applicability, and real-world relevance.

II. BACKGROUND AND RELATED WORK

[1] VisualBERT is a multimodal model that integrates a BERT-style transformer with visual features derived from object detection models like Faster R-CNN. It processes textual and visual data within a unified transformer framework, enabling the model to capture alignments between image regions and language tokens. VisualBERT has been evaluated on a range of benchmarks, including VQA 2.0, VCR, NLVR2, and Flickr30K, showcasing strong adaptability across visual-language tasks. For the VQA task, it leverages pre-training on the COCO image caption dataset to achieve competitive performance.
The training pipeline includes three stages: (1) Task-agnostic pre-training on COCO captions, (2) Task-specific pre-training using data from individual downstream tasks, and (3) Fine-tuning with task-specific objectives and additional layers to optimize performance.

Model	Datasets	Performance
VisualBERT	VQA v2.0	71.00
	NLVR2	67.00
	VCR	71.6
	Flickr30K	71.33

Table – 1: Performance table of VisualBERT with different datasets

[2] ViLBERT, a model designed for task-agnostic visiolinguistic (vision-and-language) representations. ViLBERT extends BERT to handle both visual and language inputs, enabling it to work across multiple vision-and-language tasks, such as Visual Question Answering (VQA), by pretraining on large paired datasets like Conceptual Captions. Two-Stream Architecture, ViLBERT uses separate "streams" for visual and textual data that interact through co-attentional transformer layers. This setup allows different levels of processing for each modality (visual and language) while enabling cross-modal interactions. The authors trained ViLBERT on two proxy tasks, Masked Multi-modal Learning and Multi-modal Alignment Prediction, this task requires the model

to determine if a given text correctly describes the image, enhancing the alignment between vision and language features. ViLBERT was initially pretrained on the Conceptual Captions dataset, which contains about 3.3 million image-caption pairs, and later transferred to various specific tasks, including VQA on the COCO dataset, using fine-tuning. For VQA, ViLBERT was fine-tuned to answer questions about images.

ViLBERT	VQA v2.0	70.55
	VCR	72.42

Table – 2: Performance table of ViLBERT with different datasets

[3] This paper enhances Visual Question Answering (VQA) models for autonomous driving by aligning model attention with human attention in driving scenarios. While humans prioritize critical objects like vehicles and road signs, VQA models often focus on irrelevant details, reducing accuracy in driving-related tasks. To address this misalignment, we introduce a human-guided filter that prioritizes essential driving features (e.g., lanes, traffic lights) before visual data is processed by the vision transformer. This filter is integrated into the LXMERT model's vision processing pipeline, improving interpretability, and reducing noise from irrelevant elements. Using the NuImages dataset, we evaluate the impact of this approach by comparing the original and filter-enhanced models against human responses. Results show that the filtered model better aligns with human attention patterns, leading to improved accuracy in driving-related VQA tasks. However, while this approach enhances domain-specific performance, it may limit generalization to broader VQA applications.

[4] Pixel-BERT Pixel-BERT aligns image pixels with text in an end-to-end framework, using a CNN to process pixel-level data instead of relying on region-based features from object detectors like Faster R-CNN. This preserves fine-grained visual details such as spatial relationships, object shapes, and background context. Pixel and text embeddings are combined in a multi-modal Transformer to enable rich cross-modal attention. The model is pre-trained using Masked Language Modeling (MLM) and Image-Text Matching (ITM), helping it learn contextual and relational understanding across modalities. By operating directly at the pixel level, Pixel-BERT overcomes the limitations of bounding box-based approaches.

Models	Datasets	Performance
ViLBERT	VQA v2.0	70.55
VisualBERT	VQA 2.0, NLVR2	70.80 (VQA) / 67.4 (NLVR2)
VL-BERT	VQA 2.0	72.22
LXMERT	VQA 2.0	72.54
UNITER	VQA 2.0, Flickr30K, COCO	VQA 72.27 / 63.3 (COCO)
Pixel-BERT (x152)	VQA 2.0, Flickr30K, COCO	VQA: 74.45

Table – 3: Performance table of different models with their datasets

[5] The Med-VQA paper presents the ITLTA (Image to Label to Answer) framework to tackle challenges in medical VQA, such as limited data and interpretability. Instead of using an end-to-end model, ITLTA splits the task into multi-label image classification and label-based question answering, reducing complexity and improving adaptability in data-scarce settings. A CNN-based model (DenseNet) is first pretrained on external medical datasets to classify image attributes like modality, organ, and abnormality. For answering, ITLTA uses prompting with large language models (e.g., GLM-6B, Baichuan-13B) in a zero-shot setup, avoiding end-to-end training. This design reduces reliance on annotated Med-VQA datasets while lowering computational cost.

[6] The Hi-VT5 model, introduced in the DocVQA paper, addresses Document Visual Question Answering (DocVQA) on multi-page documents using a hierarchical Transformer architecture capable of processing up to 20,480 tokens. Each page is encoded independently, producing [PAGE] tokens that summarize key information based on the question. Built on T5, Hi-VT5 concatenates these page-level embeddings and feeds them into a decoder to generate answers, effectively managing document length and structure. It is pretrained on a layout-aware de-noising task to align textual layout with semantics, enhancing the contextual value of [PAGE] tokens. A secondary module predicts the page containing the answer, improving explainability. This design reduces memory usage by summarizing pages rather than processing the full sequence at once.

[7] MMFT-BERT enhances video-based Visual Question Answering (VQA) using separate BERT-based encoders for each modality: Q-BERT for text, V-BERT for visual features, and S-BERT for subtitles. Each modality is processed independently, capturing modality-specific information before fusion. The MMFT module aggregates these outputs using a trainable [FUSE] vector that attends to relevant modalities based on the question. Multi-head attention is applied across the concatenated outputs, enabling adaptive focus on informative features. Unlike models with fixed feature extractors, MMFT-BERT is trained end-to-end, allowing dynamic learning of multimodal representations. While this improves performance in complex scenarios, the multi-stream architecture increases computational demands, posing challenges for real-time use.

Models	Performance (Q+V+S)
Two-Stream	67.70
Single BERT	72.20
WACV20	72.45
MMFT	74.97

Table – 4: Performance comparison of experiments on TVQA dataset

[8] ViLT (Vision-and-Language Transformer) removes convolutional layers and object detectors, using a patch projection technique that divides images into fixed-size patches (e.g., 32×32), linearly embeds them, and feeds them directly into a Transformer alongside text tokens. This unified architecture reduces complexity and boosts speed. ViLT is pretrained with Masked Language Modeling (MLM) and Image-Text Matching (ITM) to learn token prediction and cross-modal alignment. Its simplified design allows inference speeds up to 60× faster

than region-based models, while still achieving competitive results on standard VQA and retrieval tasks.

Models	Datasets	Performance
VilBERT	MSCOCO, Visual Genome	70.55
VisualBERT	MSCOCO, Conceptual Captions	70.80
LXMERT	MSCOCO, Visual Genome, GQA	72.42
UNITER	MSCOCO, Visual Genome, Conceptual Captions, SBU	72.70
PIXEL-BERT	MSCOCO	74.45
ViLT	MSCOCO, Visual Genome, Conceptual Captions, SBU	70.33

Table – 5: Performance comparison of different models with their datasets

[9] VBFusion introduces a multi-modal transformer-based model for Visual Question Answering (VQA) in remote sensing (RS) images, enabling natural language queries over satellite data. It uses separate encoders: a BoxExtractor for images, which generates random bounding boxes without predefined object labels, and a BERT-based tokenizer for text. Fusion occurs via a VisualBERT-based module with self-attention across image and text features, followed by an MLP for answer prediction. The model was evaluated on RSVQA-LR and RSVQAxBEN, with enhanced accuracy on the latter after including additional spectral bands, which improved performance on complex questions by enriching spatial and spectral representation.

[10] DQMA (Question-Driven Multiple Attention) enhances VQA performance by focusing on question-relevant image regions. Visual features are extracted using Faster R-CNN, while questions are embedded via GloVe and processed by an LSTM. A question-driven attention mechanism computes relevance-based attention scores to highlight essential visual areas. A co-attentive network follows, with Self-Attention (SA) capturing dependencies within the question and Guided Attention (GA) aligning text and image features. The final attended features are passed through an MLP with softmax for answer prediction. Evaluated on VQA v2.0, DQMA outperforms other models in handling complex, crowded visual scenes.

[11] VLC-BERT enhances Visual Question Answering (VQA) by integrating commonsense knowledge for tasks requiring contextual reasoning. Using Faster R-CNN for image region encoding and COMET for generating commonsense inferences, VLC-BERT dynamically incorporates contextual knowledge from ConceptNet and ATOMIC. Inferences are filtered using SBERT for semantic ranking, with the top selections fused with visual and textual features via a multi-head attention mechanism. The final fused representation is processed by a single-stream transformer (VL-BERT). Evaluated on the OK-VQA and A-OKVQA datasets, VLC-BERT outperforms models relying on static knowledge bases, demonstrating superior accuracy in tasks requiring complex reasoning.

Model	Dataset	Performance
ViLBERT	A-OKVQA	25.85
LXMERT	A-OKVQA	25.89
VLC-BERT	A-OKVQA	38.05

Table – 6: Performance table of different models with their datasets

[12] UNITER is a universal image-text representation model for various vision-and-language tasks, including VQA, image-text retrieval, and visual entailment. It uses Faster R-CNN for image region extraction and a BERT-based model for text embedding, creating joint representations through a Transformer architecture. UNITER is pretrained using Masked Language Modeling (MLM), Masked Region Modeling (MRM), and Image-Text Matching (ITM) for alignment, along with Word-Region Alignment (WRA) using Optimal Transport (OT) for fine-grained word-region mapping. Pretraining is conducted across multiple datasets (COCO, Visual Genome, Conceptual Captions, and SBU Captions), ensuring broad generalizability across V+L tasks.

Models	Performance
SOTA	70.90
VilBERT	70.92
VLBERT	72.22
VisualBERT	71.00
LXMERT	72.54
UNITER	74.02

Table – 7: Performance table of models with their datasets

III. RESEARCH GAPS AND CHALLENGES

1) Limited Utilization of Contextual Visual Features Traditional VQA models rely primarily on bounding box features, often missing fine-grained object relationships and global scene context. By integrating scene graphs, our approach enhances the model’s ability to capture and utilize object relations, spatial configurations, and attribute-level details.
2) Cross-Modal Feature Alignment Existing architectures struggle with aligning visual and textual information effectively, especially when dealing with complex scene interactions. Our approach leverages scene graphs to establish structured connections between objects and their attributes, facilitating better cross-modal reasoning and reducing ambiguities in question interpretation.
3) Efficiency in Multi-Modal Processing Multi-modal models often require high computational resources, especially when processing high-resolution images or multiple feature streams. Our scene graph-based method optimizes feature extraction by focusing on relevant object relationships rather than exhaustive pixel-level processing, improving computational efficiency without sacrificing contextual depth.

IV. METHODOLOGY

Understanding images and text together is what makes Visual Question Answering (VQA) such a challenging problem it requires reasoning at the level of human intelligence, often referred to as an AI-complete task.

Traditional VQA models, built on deep learning, typically use convolutional neural networks (CNNs) like Faster R-CNN to extract features from images, while transformers process the accompanying text. These features are then combined through a multimodal fusion mechanism to predict an answer.

However, despite advancements in transformer-based models like LXMERT, VQA systems still face major hurdles:

- Complex reasoning – Understanding spatial relationships and functional dependencies remains difficult.
- Ambiguous answers – Questions can have multiple correct answers, especially when objects share similar attributes.

To tackle these challenges, we introduce scene graphs a structured way of representing images that captures objects, their attributes, and relationships explicitly. Instead of relying solely on feature extraction, scene graphs provide a graph-based structure where:

- Objects (nodes) represent elements in the image (e.g., "dog," "table," "person").
- Attributes describe properties of these objects (e.g., "red ball," "wooden table").
- Relationships (edges) define interactions between objects (e.g., "dog chasing ball," "man sitting on chair").

By integrating scene graphs, we enhance VQA models in several ways:

- Better object-level understanding – The model does not just detect objects; it grasps their properties and relationships.
- More accurate question-answer mapping – Many VQA questions rely on understanding relationships (e.g., "Who is under the table?"), and scene graphs enable direct reasoning about such dependencies.
- Reduced ambiguity – Explicitly encoding relationships minimizes misinterpretation of complex scenes.
- Improved generalization – Instead of memorizing dataset-specific patterns, the model learns reusable object-relation structures, making it more adaptable to unseen questions.
- Less dependency on massive datasets – Traditional models need large amounts of data to learn implicit relationships, while scene graphs provide this knowledge explicitly, leading to more efficient learning.
- More efficient attention mechanisms – Instead of analysing the entire image, the model can focus on relevant objects and their interactions, reducing computational overhead and improving accuracy.

A. *Datasets Used*

1) VQA v2.0 Dataset

Contains image-question-answer triplets, where the goal is to answer questions based on image content.

However, it does not provide structured relational information, making complex reasoning difficult.

Preprocessing:

a) Image Preprocessing:

- Images were processed using the Faster R-CNN model to extract visual features.
- The extracted features were stored in a PKL (Pickle) file for efficient storage and retrieval.

b) Combining Processed Data:

- The pre-processed image features (from the PKL file) were combined with the questions and answers dataset.
- A final PKL file was created, containing the following fields:

Answers	List of 10 human annotated answers
Question ID	Unique identifier for each question
Image ID	Unique identifier for each image
Question	The natural language question about the image
Image Features	Extracted image feature vectors from Faster R-CNN Array of shape (N, 4), N varies

Table – 8: VQA v2.0 PKL file attributes

Total Entries	Number of QA pairs per image 3-5 Number of unique images in Train: 82783 Validation: 40504 Total: Training: 443744 Validation: 214336

Table – 9: VQA v2.0

2) Visual Genome (VG) Dataset

Contains detailed scene graphs for images.

Serves as an auxiliary dataset to train the model to understand explicit relationships.

Preprocessing:

a) Image Preprocessing:

- Followed the same process as the VQA v2.0 dataset, where images were processed using Faster R-CNN and stored in a PKL file.

Answers	1 human annotated answer
Question ID	Unique identifier for each question
Image ID	Unique identifier for each image
Question	The natural language question about the image
Image Features	Extracted image feature vectors from Faster R-CNN Array of shape (N, 4), N varies

Table – 10: Visual Genome PKL file attributes

Total Entries	Number of QA pairs per image ~17 avg Number of unique images in Train: 64346 Validation: 43903 Total Used: Training: 443744 Validation: 214336

Table – 11: Visual Genome

b) Scene Graph Data Processing:

- Relationships and attributes were extracted from the dataset and stored as JSON files.

c) Processing During Training (Data Loader):

- Instead of preprocessing scene graph data beforehand, relationships and attributes were processed dynamically in the data loader.
- Extracted the following graph components:
    - Nodes: Objects detected in the image.
    - Edges: Relationships between objects.
    - Edge Attributes: Additional attributes defining the edges (e.g., spatial relations)

JSON File	Key Fields
Attributes	Image id: ID of the image objects: List of objects with attributes object id: Unique ID for each object names: List of object names attributes: List of attribute labels for the object Bounding box (coordinates defining object location)
Relationships	Image id: ID of the image relationships: List of relationships subject: Object initiating the relationship predicate: Relationship type (e.g., "on", "next to")

	object: Object receiving the relationship Bounding boxes for both subject and object

Table – 12: Scene Graph

B. *Architecture*

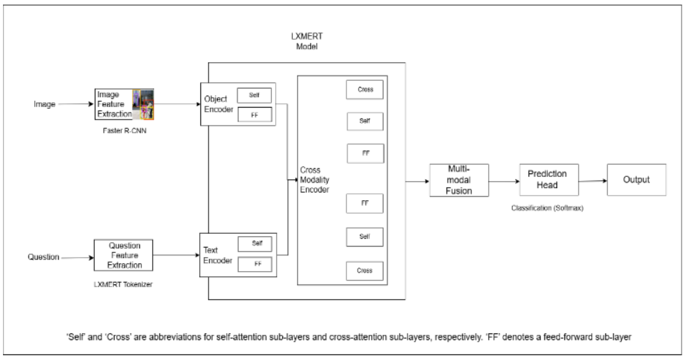

Fig – 1: Baseline LXMERT Model for VQA

Key Components:

1) Image Feature Extraction
   - Uses Faster R-CNN to extract visual object features from the image.
2) Question Feature Extraction
   - Uses BERT Tokenizer to convert the question into a feature representation.
3) LXMERT Model: Encoding & Fusion
   - Object Encoder: Processes extracted image features using self-attention (Self) and feed-forward (FF) layers.
   - Text Encoder: Processes the tokenized question using self-attention (Self) and feed-forward (FF) layers.
   - Cross-Modality Encoder: Enables interaction between the image and question features through cross-attention and self-attention mechanisms.
4) Prediction Module
   - The Multi-Modal Fusion layer integrates vision and text features.
   - A classification head (Softmax layer) predicts the final answer.

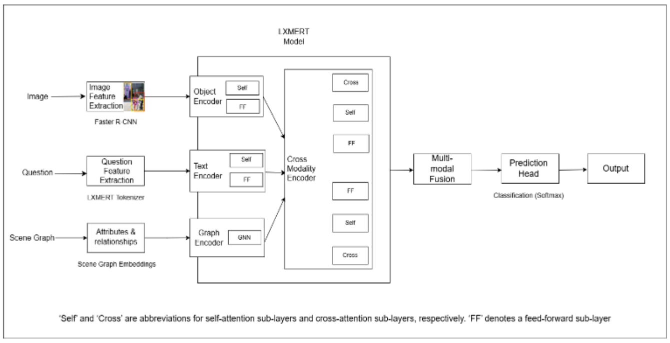

Fig – 2: LXMERT Model with Scene Graphs

Components

1) Scene Graph Embeddings:
   - Extracts attributes (e.g., "red", "large") and relationships (e.g., "man sitting on chair").
2) Graph Neural Network (GNN) Encoder:
   - Processes scene graph embeddings using a Graph Neural Network (GNN).
   - Encodes spatial, semantic, and functional relationships between objects.
3) Integration into LXMERT:
   - The Graph Encoder outputs relational features, which are fused with the Image and Text.
4) Prediction Module (Same as Baseline LXMERT)
   - The multi-modal fusion combines image, text, and scene graph features.
   - The classification head outputs the final prediction.

C. *Implementation*

1) *Model Combination Strategy*

Confidence-Based Model Selection (CBMS) for VQA & VG

How It Works:

a) Compute confidence scores for both models:
   - Each model outputs logits (raw prediction scores before SoftMax).
   - The confidence score is computed as the SoftMax probability of the predicted answer.

$$C_{\mathrm{VQA}} = \max\left(\mathrm{softmax}\left(logits_{\mathrm{VQA}}\right)\right.$$

$$C_{\mathrm{SG}} = \max\left(\mathrm{softmax}\left(logits_{\mathrm{SG}}\right)\right)$$

b) Select the more confident model for final prediction:
   - If $C_{VQA}>C_{SG}$, use the VQA model's answer.
   - Otherwise, use the Scene Graph model's answer.

$$P_{\mathrm{final}} = P_{\mathrm{VQA}} \cdot \mathbb{1}\left(C_{\mathrm{VQA}} > C_{\mathrm{SG}}\right) + P_{\mathrm{SG}} \cdot \mathbb{1}\left(C_{\mathrm{SG}} \geq C_{\mathrm{VQA}}\right)$$

- Why This Approach?
  - The VQA model performs better on direct visual questions (e.g., "What colour is the car?").
  - The Scene Graph model is superior for relational queries (e.g., "Who is sitting next to the woman?").
  - By dynamically selecting the more confident model per question, we get the best of both models.

2) *Training Process*

The model is trained in two phases before combining them using CBMS.

a) Training the VQA Model (LXMERT)

- Dataset: VQA v2.0
- Loss Function: Cross-Entropy Loss
- Optimizer: Adam
- Batch Size: 32
- Epochs: 10

b) Training the Scene Graph Model (SG-VQA)

- Dataset: Visual Genome Scene Graphs
- GNN Architecture:
  - Node embeddings: 128-dim features (GloVe embeddings)
  - Edge embeddings: 64-dim features
  - Aggregation: Graph Attention Networks (GAT)
- Loss Function: Cross-Entropy Loss

- Optimizer: Adam
- Batch Size: 32
- Epochs: 10

V. RESULTS

Performance Comparison Across Datasets

To evaluate the effectiveness of scene graphs, we conducted experiments using the VQA v2.0 dataset and the Visual Genome (VG) dataset. The results are summarized in the table below:

Dataset	Input	Training
VQAv2.0	Q + A + I	28%
VG	Q + A + I	29%
VG + SG	Q + A + I + A + R	39%

Table – 13: Individual Model Results

A. Observations:

1) Baseline Performance (VQA v2.0 & VG)
   - Training accuracy for VQA v2.0 and Visual Genome remained close (~28-29%), showing similar learning capacity.
   - Validation accuracy was lower (24-25%), indicating overfitting on training data.
2) Impact of Scene Graphs (VG + Scene Graph)
   - Training accuracy jumped to 39%, proving that structured relationships help the model learn object dependencies better.
   - However, validation accuracy remained similar (23%), suggesting a need for better generalization strategies.

B. Final Model Performance (VQA + Scene Graph Model)

After implementing Confidence-Based Model Selection (CBMS) to combine the VQA v2.0-trained model and the VG Scene Graph-based model, the accuracy improved significantly:

Dataset	Accuracy
VQAv2.0 Sample1k	~43%
Visual Genome Sample1k	~59%

Table – 14: Combined Model Results

C. Key Takeaways

Accuracy Improvement

- The final VQA + Scene Graph model outperformed the standalone models on both datasets.
- VQA v2.0 accuracy increased 43%, proving that scene graphs enhance reasoning capabilities.
- Visual Genome accuracy improved significantly (59%), showing that scene graphs are highly effective in datasets rich in relational information.

Generalization & Challenges

- While accuracy increased, validation accuracy did not improve drastically (23%), indicating potential domain shift issues.
- This suggests that while scene graphs help in understanding relationships, there is a need for better fine-tuning strategies for diverse question types.

VI. CONCLUSION AND FUTURE SCOPE

This study enhances Visual Question Answering (VQA) by integrating scene graphs, enabling the model to capture fine-grained object relationships and contextual dependencies beyond bounding box features. By improving cross-modal reasoning and reducing computational inefficiencies, our approach addresses key challenges in traditional VQA systems. The results demonstrate that leveraging structured visual representations enhances interpretability and accuracy, making the model more effective in complex question-answering scenarios.

Future work will focus on expanding the system to handle domain-specific queries across diverse datasets, such as medical imaging, autonomous driving, and satellite imagery. Incorporating heterogeneous datasets would improve the model's adaptability to real-world applications. Additionally, optimizing the model for large-scale deployment using cloud-based infrastructures with distributed GPUs or TPUs would enable real-time processing. These advancements could transform VQA into a scalable AI service for industries requiring advanced visual understanding, such as healthcare, defence, and smart city applications.

REFERENCES

[1] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J. and Chang, K.W., 2019. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

[2] 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.

[3] Rekanar, K., Hayes, M., Sistu, G. and Eising, C., 2024. Optimizing visual question answering models for driving: Bridging the gap between human and machine attention patterns. arXiv preprint arXiv:2406.09203.

[4] Huang, Z., Zeng, Z., Liu, B., Fu, D. and Fu, J., 2020. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849.

[5] Wang, Jianfeng, Seng, Kah, Shen, Yi, Ang, Li-Minn, and Huang, Difeng, 2024. Image to label to answer: An efficient framework for enhanced clinical applications in medical visual question answering. Electronics, 13, 2273.

[6] Tito, R., Karatzas, D. and Valveny, E., 2023. Hierarchical multimodal transformers for multipage DocVQA. Pattern Recognition, 144, p.109834.

[7] Khan, A.U., Mazaheri, A., Lobo, N.D.V. and Shah, M., 2020. MMFT-BERT: Multimodal fusion transformer with BERT encodings for visual question answering. arXiv preprint arXiv:2010.14095.

[8] 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.

[9] Siebert, T., Clasen, K.N., Ravanbakhsh, M. and Demir, B., 2022, October. Multimodal fusion transformer for visual question answering in remote sensing. In Image and Signal Processing for Remote Sensing XXVIII (Vol. 12267, pp. 162-170). SPIE.

[10] Wu, J., Ge, F., Shu, P., Ma, L. and Hao, Y., 2022. Question-driven multiple attention (DQMA) model for visual question answer. 2022 International Conference on Artificial Intelligence and Computer Information Technology (AICIT), Yichang, China, pp. 1-4. https://doi.org/10.1109/AICIT55386.2022.9930294.

[11] Ravi, S., Chinchure, A., Sigal, L., Liao, R., and Shwartz, V., 2023. VLC-BERT: Visual question answering with contextualized commonsense knowledge. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 1155-1165. https://doi.org/10.1109/WACV56688.2023.00121.

[12] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J., 2020, August. UNITER: Universal image-text representation learning. In European Conference on Computer Vision (pp. 104-120). Cham: Springer International Publishing.