Visual Question Answering with Scene Graphs

Saurabh Madake
Department of Computer Science and Engineering
COEP Technological University (COEP Tech)
Pune, India - 411005
madakesb23.comp@coeptech.ac.in

Archana Patil
Department of Computer Science and Engineering
COEP Technological University (COEP Tech)
Pune, India - 411005
abp.comp@coeptech.ac.in

Aparna Santra Biswas
Department of Computer Science and Engineering
COEP Technological University (COEP Tech)
Pune, India - 411005
aparna.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. In addition to the standard model, we propose a Confidence-Based Model Selection (CBMS) fusion strategy to improve prediction reliability. By computing the SoftMax confidence of logits from both the VQA and Visual Genome trained models, the system selects the model with higher confidence for each prediction. This fusion approach achieves a consensus accuracy of 65.01%. 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, Confidence-Based Model Selection, Multi-Modal Deep Learning

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. 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 Visual Question Answering (VQA). Traditional approaches relied on separate processing pipelines for visual and textual data, often leading to integration challenges. In contrast, transformer architectures such as LXMERT process multimodal data simultaneously, significantly improving VQA accuracy and performance. By leveraging cross-modal attention mechanisms, LXMERT enables deep alignment between visual content and language, making it highly effective for this task.

While LXMERT has demonstrated strong performance on VQA benchmarks, our work builds upon and extends this foundation in several meaningful ways. Specifically, we propose a modified VQA pipeline that not only evaluates LXMERT on standard datasets like VQA v2.0 and Visual Genome, but also incorporates scene graph-based relational reasoning from Visual Genome using a graph neural network (GNN) module. This enhancement allows the model to better capture fine-grained relationships and attributes among objects in an image information that traditional LXMERT pipelines often underutilize.

Furthermore, we introduce a comparative analysis framework that evaluates baseline LXMERT performance against our enhanced model, highlighting how the incorporation of structured scene representations improves reasoning in complex scenarios. Implementing both baseline and modified architectures within robust frameworks such as PyTorch and Hugging Face facilitates an in-depth exploration of the model's capabilities, limitations, and areas for improvement.

Through this study, we not only evaluate LXMERT's effectiveness in handling VQA tasks but also demonstrate how enriching multimodal transformers with external visual reasoning modules can lead to measurable performance gains. By refining transformer-based methodologies, our research contributes to advancing the field of multimodal AI, enhancing model accuracy, interpretability, and real-world applicability.

II. BACKGROUND AND RELATED WORK

[1] VisualBERT integrates visual region embeddings directly into a single BERT-based transformer by concatenating them with textual tokens. This early fusion method allows tight integration of modalities from the start, simplifying architecture but potentially diluting modality-specific nuances [1].

[2] ViLBERT introduces a two-stream architecture where visual and textual streams are processed separately but interact via co-attentional transformer layers. Unlike LXMERT, it processes the streams in parallel for longer periods before cross-modal fusion, offering modularity and generality across tasks [2].

[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, 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 [3].

[4] Pixel-BERT abandons region-based input in favour of pixel-level CNN features, which are then fused with text in a shared transformer encoder. This direct pixel approach allows capturing fine spatial information but increases computational load [4].

[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 [5].

[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 [6].

[7] MMFT-BERT separates modalities into independent BERT encoders (text, vision, subtitles) and uses a [FUSE] vector to perform adaptive cross-modal fusion, enabling modality-aware reasoning particularly effective for video-based VQA [7].

[8] ViLT removes the need for object detectors entirely by employing patch embeddings for images and fusing them with text tokens in a pure transformer model. This significantly reduces inference time, making it a lightweight alternative at the cost of some spatial precision [8].

[9] VBFusion employs box-based visual encoders without labelled objects and combines them with BERT text encoders. The fusion mechanism is modelled on VisualBERT, with enhancements for spatial-spectral reasoning in remote sensing images [9].

[10] DQMA (Question-Driven Multiple Attention) adds guided and self-attention modules tailored to question relevance, emphasizing regions critical to the query. It relies on traditional LSTM encodings and object detection but introduces a more task-specific fusion mechanism [10].

[11] VLC-BERT enhances standard transformer pipelines by incorporating commonsense knowledge through dynamic retrieval from external sources (e.g., ConceptNet), which is then merged with visual and textual streams using multi-head attention fusion [11].

[12] UNITER uses a unified transformer encoder with shared processing of both modalities and introduces Word-Region Alignment (WRA) via Optimal Transport, allowing for highly precise vision-language matching [12].

Models	Datasets	Accuracy
VisualBERT [1]	VQA v2.0	71.00
	NLVR2	67.00
	VCR	71.60
	Flickr30K	71.33
ViLBERT [2]	VQA v2.0	70.55
	VCR	72.42
Pixel-BERT (x152) [4]	VQA 2.0	74.45
MMFT [7]	TVQA	74.97
ViLT [8]	Visual Genome	70.33
VLC-BERT [11]	A-OKVQA	38.05
UNITER [12]	COCO, Visual Genome, Conceptual Captions, and SBU Captions	74.02

Table - 1: Accuracy Table of VQA models with their Datasets

III. RESEARCH GAPS AND CHALLENGES

1) We acknowledge the potential data leakage due to overlapping COCO images between Visual Genome and VQA v2.0. However, due to inconsistent image ID conventions and lack of shared file metadata, we were unable to reliably identify and filter overlapping samples. We plan to explore hash-based or visual-feature-based deduplication in future work.

2) 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.

3) 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.

4) 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.

Comparative Analysis of VQA Model Architectures

Model	Architecture Type	Visual Input	Fusion Mechanism	Unique Feature
LXMERT	Dual-Stream + Cross-Modality	Object Regions (FRCNN)	Cross-Modality Attention	Separate encoders + deep cross-modal fusion
VisualBERT [1]	Single-Stream	Object Regions (FRCNN)	Early Fusion in BERT	Simplified integration
ViLBERT [2]	Two-Stream	Object Regions (FRCNN)	Co-attentional Transformer	Parallel streams before fusion
Pixel-BERT [4]	Single-Stream	Pixel-level CNN	Shared Transformer	Fine-grained spatial encoding
MMFT-BERT [7]	Multi-stream	Region + Subtitles	Modality-Adaptive [FUSE] Vector	Modality-specific BERTs + attention fusion
ViLT [8]	Single-Stream	Patch Embeddings	Unified Transformer	No object detector, lightweight
VBFusion [9]	Hybrid	Box Proposals (Unlabelled)	Self-Attention in VisualBERT	Remote sensing focus, no object labels
DQMA [10]	Guided Attention	Object Regions (FRCNN)	Guided + Self Attention	Question-driven region focus
VLC-BERT [11]	Single-Stream	Object Regions (FRCNN)	Multi-head Attention	Dynamic commonsense knowledge
UNITER [12]	Single-Stream	Object Regions (FRCNN)	Shared Transformer + WRA	Precise word-region alignment

Table - 2: Comparative Analysis Table of VQA Model Architectures

IV. METHODOLOGY

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 - 3: 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 Used: Training: 50000 Validation: 25000

Table - 4: 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 - 5: 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: 50000 Validation: 25000

Table - 6: 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 - 7: Scene Graph

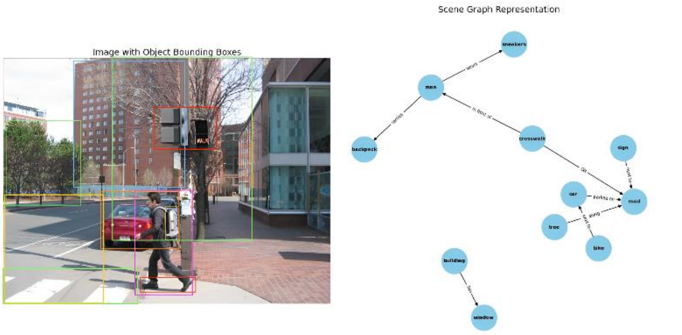
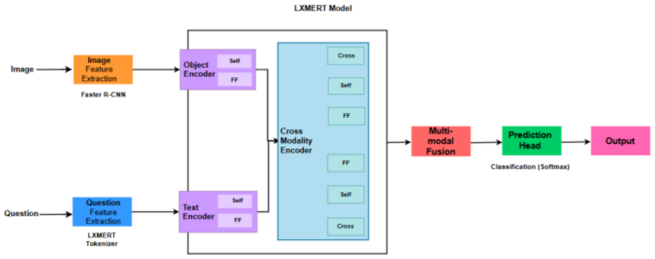


Figure - 1: Scene Graph Representation

B. *Architecture*



'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer

Figure - 2: Architecture of the baseline LXMERT model, showing image and text feature extraction followed by cross-modal fusion

Key Components:

1) Image Feature Extraction

- Uses Faster R-CNN to extract visual object features from the image.

2) Question Feature Extraction

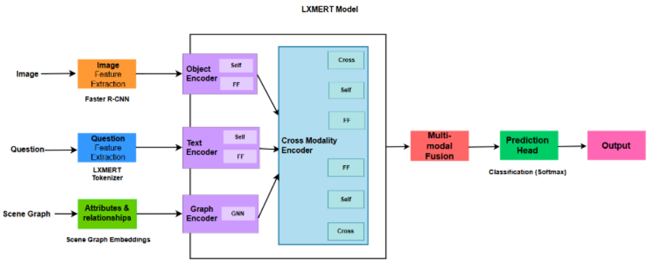
- Uses BERT Tokenizer to convert the question into a feature representation.

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.



'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer, 'GNN' denotes Graph Neural Network

Figure - 3: Proposed LXMERT model with scene graph integration via a GNN encoder, enhancing relational reasoning

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.

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

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

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

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

- 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

3) *Addressing Overfitting*

To improve generalization and reduce overfitting, we used a validation set during training, skipped ambiguous samples, and incorporated an <unk> token in the loss function to handle rare answers. Gradient accumulation (with 4 steps) further stabilized training by simulating a larger batch size without extra memory overhead

V. RESULTS

Performance Comparison Across Datasets

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

Dataset	Input	Training Accuracy
VQAv2.0	Q + A + I	37%
VG + SG	Q + A + I + A + R	45%

Table - 7: Individual Model Results

A. *Observations:*

1) Baseline Performance - VQA v2.0
   - The model trained on the VQA v2.0 dataset (Q + A + I) achieved a training accuracy of 37%.
   - The validation accuracy was 30%, indicating a modest gap between training and generalization.

2) Enhanced Input with Scene Graphs - VG + SG
   - The model trained on the Visual Genome dataset with Scene Graphs (Q + A + I + A + R) reached a training accuracy of 45%.
   - Validation accuracy improved to 35%, showing better generalization than the VQA-only model.
   - This supports the idea that structured inputs like attributes and relationships (scene graphs) help the model learn object dependencies and contextual relationships more effectively.

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:

Fusion Strategies	Accuracy	Consensus
CBMS	59.80%	65.01%
Majority Vote Fusion	56.50%	62.03%
Avg Logits Fusion	53.80%	59.50%

Table - 8: Combined Model Results on 1k VQAv2.0 Samples

C. Key Takeaways:

1) CBMS (Confidence-Based Model Switching) performs best overall
    a) Achieves the highest hard accuracy (59.80%) and consensus (soft) accuracy (65.01%).
    b) Demonstrates that selecting predictions based on model confidence allows more reliable decision-making than simple averaging or voting.
2) Majority Vote Fusion is a strong baseline
    a) Performs better than average logits fusion with 56.50% accuracy and 62.03% consensus.
    b) This indicates that agreement among models can be a useful heuristic, even without considering confidence scores.
3) Averaging Logits (Soft Voting) performs the worst
    a) Produces the lowest accuracy (53.80%) and consensus (59.50%).
    b) This suggests that the logits from both models may not be well-calibrated or aligned, and averaging them dilutes model-specific strengths.
4) Confidence-based fusion strategies are more effective
    a) Incorporating model certainty (confidence) into the fusion decision leads to more accurate and human-aligned predictions.
    b) CBMS intelligently leverages the strengths of both models without forcing agreement.

VI. CONCLUSION AND FUTURE SCOPE

In this work, we proposed a Confidence-Based Model Switching (CBMS) strategy to fuse predictions from two separately trained LXMERT models one on the VQA v2.0 dataset and the other on Visual Genome with scene graphs. Our evaluation demonstrates that CBMS outperforms traditional fusion methods such as majority voting and average logits, achieving the highest hard accuracy (59.80%) and VQA-standard consensus accuracy (65.01%). Our ablation studies further confirmed the benefits of structured input, with the VG + Scene Graph model achieving 45% training accuracy, surpassing the 37% baseline from VQA alone. Additionally, we incorporated soft accuracy metrics to reflect human consensus and analysed failure cases to understand model limitations. In future work, we aim to extend fusion techniques using learnable attention-based or confidence-weighted ensembles, explore visual similarity-based de-duplication for more robust overlap filtering, and scale experiments using larger subsets of scene graph annotations to further enhance reasoning capabilities.

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.