Transformer-Driven Visual Question Answering on Complex Datasets

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

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

Abstract **- The project focuses on the development of 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 the system to answer questions based on the content of an input image, making it a challenging problem in the field of computer vision and natural language processing. The project 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 project contributes to the field by pushing the boundaries of multi-modal deep learning and explores the potential of transformers in addressing complex, AI-complete tasks.**

Key Words: Transformers, VQA, LXMERT, Visual Genome

1.INTRODUCTION

The project on Visual Question Answering (VQA) represents a cutting-edge exploration in the fields of computer vision and natural language processing, aiming to bridge the gap between visual and textual understanding. As an AI-complete task, VQA requires a high level of computational understanding, making it a significant challenge in artificial intelligence research. The core of VQA involves developing systems capable of interpreting visual content and responding accurately to questions about it, with potential applications ranging from aiding visually impaired individuals to enhancing human-computer interactions. By enabling machines to answer questions based on visual inputs, VQA pushes the boundaries of machine comprehension and has profound implications for real-world, assistive technologies.

Project focuses specifically on the use of transformer-based models, such as LXMERT, to tackle the complexities of VQA. Traditional approaches in this field often relied on separate processing pipelines for visual and textual data, which posed challenges for integrating information seamlessly. However, transformer models like LXMERT have brought transformative advancements by processing multimodal data simultaneously, significantly improving VQA accuracy and performance. Leveraging the attention mechanisms in transformers, LXMERT enables deep alignment between visual content and language, making it highly suitable for VQA tasks. By integrating LXMERT with robust frameworks such as PyTorch and Hugging Face, this project sets the stage for exploring the model's potential and limitations within this complex domain.

To support analysis part, this project utilizes well-established datasets like VQA v2.0 and Visual Genome, which provide diverse images, questions, and associated metadata. These datasets serve as valuable resources for training and evaluating the model, as they offer varied scenes, question types, and visual relationships critical for robust VQA performance. Through this project, we aim not only to evaluate the capabilities of LXMERT in handling VQA but also to provide insights into the evolution, current state, and future directions of VQA technology. This comprehensive approach positions this project at the forefront of VQA research, with a focus on improving model accuracy, applicability, and practical relevance in multimodal AI.

2. BACKGROUND AND RELATED WORK

[1] VisualBERT combines a BERT-like transformer with visual embeddings derived from object proposals (using tools like Faster R-CNN). Both text and image data are processed in a unified Transformer structure, allowing the model to learn associations between image regions and language tokens. VisualBERT was tested on VQA 2.0 for question answering on images, VCR for commonsense reasoning in visual contexts, NLVR2 for visual-language reasoning, and Flickr30K for grounding phrases to image regions. VisualBERT showed high adaptability across diverse tasks. For VQA, it used COCO pre-training to achieve competitive results. Unlike my approach (LXMERT), VisualBERT emphasizes a more straightforward integration of visual and textual inputs, showing benefits in efficiency and interpretability, though potentially lacking in some nuanced multimodal interactions that LXMERT's split-transformer model offers.
Training:
Task-Agnostic Pre-training: The model is initially pre-trained on the COCO image caption dataset.
Task-Specific Pre-training: Before fine-tuning on specific tasks, the model undergoes a second pre-training phase on data from each target task.

Fine-Tuning: The model is fine-tuned on each task with additional layers and objectives tailored for performance.

[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. The authors observed that ViLBERT's pretraining improved its performance on VQA and other tasks over task-specific models, suggesting strong transferability of its visiolinguistic representations.

[3] The paper focuses on enhancing Visual Question Answering (VQA) models for autonomous driving by aligning the attention patterns of the model with human attention in driving contexts. The authors observe a gap between human attention patterns and VQA model outputs in driving scenarios. While humans focus on objects critical for driving (e.g., vehicles, road signs), models often attend to irrelevant details like trees or buildings. This misalignment can reduce the model's effectiveness in answering driving-related questions accurately. To bridge this gap, the authors integrate a human-guided "filter" that prioritizes relevant driving features (e.g., lanes, traffic lights) before passing visual data through the vision transformer. This filter ensures that the model's attention mechanism aligns more closely with what a human would focus on in driving scenarios. They modify the LXMERT model, which is structured with separate encoders for vision, language, and cross-modality processing, by inserting this filter into the vision processing pipeline. The filtered data helps the model focus on essential driving-related features, improving interpretability and reducing computational noise from irrelevant features. Using the NuImages dataset, they evaluate the filter's impact on LXMERT's performance by comparing outputs from the original and filter-enhanced models against human responses. The results show that the filtered model aligns more closely with human attention patterns, improving the model's accuracy on driving-related VQA tasks. While this can enhance domain-specific performance, it might limit generalization to other VQA contexts.

[4] Pixel-BERT, a model for aligning image pixels directly with text in a unified, end-to-end framework. Instead of using object detection models like Faster R-CNN for pre-extracting region-based features (bounding boxes), Pixel-BERT uses a convolutional neural network (CNN) to directly process pixel-level image data. This approach retains more fine-grained visual information, including spatial relationships, object shapes, and background details that region-based methods might overlook. Pixel-level embedding is integrated with the text embeddings in a multi-modal Transformer to facilitate rich cross-modal attention. To improve the robustness of visual representation and prevent overfitting, Pixel-BERT randomly samples pixels during pre-training, promoting flexibility and generalization in learning from incomplete visual inputs. This method reduces the computational load while enhancing the ability to capture diverse visual patterns. The model is pre-trained with Masked Language Modeling (MLM) and Image-Text Matching (ITM), similar to other multi-modal models. In MLM, Pixel-BERT learns to predict masked words using contextual clues from both visual and textual data. ITM involves distinguishing paired image-text pairs from randomly shuffled ones, improving the model's understanding of cross-modal relationships. The architecture of Pixel-BERT allows end-to-end learning of both CNN and Transformer layers. Unlike models with pre-computed region features, this setup enables the model to adapt its visual representations dynamically to each task, breaking the limitation of predefined detection categories. By working directly at the pixel level, Pixel-BERT avoids the limitations of bounding boxes, such as missing details on shapes and spatial relationships between objects. Pixel-level processing can be computationally intensive, particularly with high-resolution images, as it processes all pixels rather than focusing on a limited number of regions. LXMERT is well-suited for tasks with distinct, identifiable objects (e.g., objects in driving scenes), while Pixel-BERT may perform better in scenes where holistic visual context is important, such as complex spatial queries.

[5] The paper on Med-VQA introduces the ITLTA (Image to Label to Answer) framework, tailored to address challenges in Medical Visual Question Answering (Med-VQA) by efficiently handling limited data and enhancing interpretability. Unlike end-to-end VQA models, ITLTA decomposes the task into two main parts: multi-label learning for medical images and label-based question answering. This approach reduces complexity, allowing it to adapt to scarce data environments typical in medical applications. ITLTA first pretrains a CNN-based model (Densenet) using external medical image datasets to classify image attributes like modality, organ, and abnormality, crucial for handling limited annotated data in Med-VQA. It employs multi-task learning for efficient pretraining and integrates a joint feature and label

attention mechanism. This allows cross-modal correlation learning between image features and label embeddings, enhancing the model's ability to identify medically relevant image attributes. ITLTA uses a “prompting” technique to employ large language models (LLMs) like GLM-6B or Baichuan-13B to answer questions. Instead of training a language model end-to-end, ITLTA relies on the language model’s zero-shot capabilities, minimizing data requirements and computational cost. By separately generating image labels and feeding them into the LLM alongside the question, ITLTA improves interpretability. Errors can be traced to either visual label inaccuracies or misinterpretations by the LLM, offering clearer insights into the decision-making process. By using pretraining on external datasets and leveraging zero-shot learning in LLMs, ITLTA minimizes dependency on large annotated Med-VQA datasets. While LXMERT is pre-trained on general VQA datasets and optimized for robust cross-modal embeddings, ITLTA’s staged approach leverages pretraining and zero-shot capabilities for efficiency in specialized domains with limited data.

[6] The DocVQA paper introduces Hi-VT5, a model designed to address the challenges of Document Visual Question Answering (DocVQA) on multi-page documents. Hi-VT5 handles multi-page documents by employing a hierarchical Transformer architecture that can process a sequence up to 20,480 tokens sufficient for multi-page setups in the MP-DocVQA dataset. The model’s encoder processes each page independently, generating [PAGE] tokens that summarize the essential information of each page based on the question context. Hi-VT5 builds on T5, with an encoder that generates page-level embeddings. These embeddings are concatenated across pages and fed into a decoder to generate answers, effectively distilling relevant information from large documents while handling the page structure. This hierarchical approach reduces computational complexity by processing each page separately, then aggregating page summaries for final answer generation. The model is pretrained on a hierarchical layout-aware de-noising task, which helps align textual layout with semantic features, thereby preparing the [PAGE] tokens to encapsulate contextual information from each page. This aids in better answering multi-page document queries. Hi-VT5 includes a secondary module to predict the page containing the answer. This page prediction serves as an explainability feature, offering insights into where the model found its response within a multi-page document. With this Hi-VT5 can handle long documents without needing full sequence processing, as it generates compact summaries per page, which reduces memory and computational requirements. In comparison LXMERT processes visual and text embeddings in an end-to-end model, Hi-VT5 uses hierarchical summarization, which makes it more efficient for handling long documents.

[7] The paper on MMFT-BERT (Multimodal Fusion Transformer with BERT Encodings) presents a model designed to enhance video-based Visual Question Answering (VQA) by leveraging separate BERT-based encoders for each modality (text, video, and subtitles). MMFT-BERT uses individual BERT encoders for each modality: Q-BERT for text (question-answer pair), V-BERT for visual features (from detected objects in the video), and S-BERT for subtitles. Each encoder processes its modality independently, allowing the model to capture modality-specific information before combining it. This approach enables each stream to focus on answering questions related to its specific modality, which can enhance understanding when a question relies heavily on either visual, text, or subtitle information. The MMFT module combines information from the three BERT streams (Q-BERT, V-BERT, and S-BERT) by aggregating their outputs. It uses a trainable vector called [FUSE], which attends to each modality based on relevance to the given question. The fusion process allows the model to adaptively emphasize certain modalities depending on the question's nature. MMFT applies multi-head attention across the concatenated outputs, enhancing the joint encoding by learning which parts of each modality are most informative for generating the final answer. MMFT-BERT’s objective function combines individual loss terms for each BERT encoder (text, video, subtitles) alongside a joint loss over the aggregated multimodal representation. This hierarchical loss helps the model learn both the individual importance of each modality and how they contribute to answering questions collectively. Unlike many VQA models, which rely on fixed feature extractors, MMFT-BERT’s architecture is trained end-to-end. This means the encoders and fusion module are optimized together, helping the model learn a robust multimodal representation dynamically rather than relying on static features. By allowing separate streams for each modality, MMFT-BERT improves handling of questions specific to visual content or subtitles, making it more flexible and effective in capturing relevant details. The attention mechanism in the MMFT module adapts to the question type, focusing on the most relevant modality. This leads to more accurate answers, particularly in complex, multimodal contexts. But the model's reliance on multiple BERT streams and a separate fusion transformer increases its complexity and resource requirements, particularly for real-time applications.

[8] The ViLT (Vision-and-Language Transformer) paper proposes a vision-and-language model that differs from traditional approaches by eliminating convolutional layers and object detectors for image processing. ViLT uses a patch projection technique where images are divided into fixed-size patches (e.g., 32x32 pixels). Each patch is linearly embedded, similar to how text tokens are embedded, and passed directly to the Transformer model. This design eliminates the need for complex visual processing modules, such as convolutional networks or region-based object detectors, making ViLT significantly faster than models that rely on pre-extracted region features. ViLT processes text and image embeddings

together within a single-stream Transformer. Unlike dual-stream models like LXMERT, which maintains separate streams for each modality, ViLT treats both visual and textual tokens within the same architecture, which reduces model complexity and computational load. The model is pretrained using Masked Language Modeling (MLM) and Image-Text Matching (ITM). These tasks help the model learn to predict masked tokens within text and distinguish aligned versus misaligned image-text pairs. ViLT also incorporates a Whole Word Masking (WWM) strategy and image augmentations (such as RandAugment), which enhance the model's performance on downstream tasks. With its simplified structure, ViLT achieves inference speeds up to 60 times faster than region-based models while maintaining competitive performance across standard VQA and retrieval benchmarks. By removing convolutional layers and complex region-based processing, ViLT reduces computational requirements, enabling faster training and inference times. Although without a region-based approach, ViLT may lack the fine-grained detail that object detectors capture, potentially impacting performance on tasks that rely on specific object recognition.

[9] The paper discusses a novel approach to Visual Question Answering (VQA) in remote sensing images using a multi-modal transformer-based model called VBFusion. It addresses the need for an efficient way to interpret remote sensing (RS) images by using VQA, where users can query images directly in natural language. Their method is geared towards the fusion of multi-modal data images and text questions using a joint representation approach rather than traditional methods that use modality-specific representations. They use two separate encoders for images and text. For images, a BoxExtractor randomly generates bounding boxes to identify regions of interest without predefined object labels, making it less constrained by object-based data. These boxes are processed by a ResNet-based encoder. For text, a BERT-based tokenizer processes the questions. The fusion occurs in a VisualBERT-based module with multiple transformer layers that enable self-attention across both image and text features. This joint representation aids in capturing complex relationships across modalities. The fused data is passed through a Multi-Layer Perceptron (MLP) to generate an output answer. The model was tested on two datasets RSVQA-LR (a smaller dataset of RGB bands from Sentinel-2 images) and RSVQAxBEN (a larger dataset also based on Sentinel-2 images but extended to all spectral bands). They further modified RSVQAxBEN by including more spectral bands, which enhanced the model's accuracy by providing richer spatial and spectral data. The VBFusion model learns a joint representation, enabling it to better understand image-question relationships compared to methods that only concatenate features. Including additional spectral bands notably improved performance for complex questions. VBFusion is adapted specifically for remote sensing and takes advantage of spectral bands, making it less applicable to general-purpose VQA.

[10] The paper describes a Visual Question Answering (VQA) model called the Question-Driven Multiple Attention (DQMA) model. The DQMA model aims to address challenges in existing VQA models, which often capture irrelevant details in complex scenes, leading to lower accuracy in answering questions. By focusing attention on question-relevant regions, DQMA improves model accuracy on tasks requiring selective focus. Visual features are extracted using Faster R-CNN, which detects objects and salient regions in an image, providing object proposals as visual inputs. For text, questions are embedded using Glove embeddings and processed by an LSTM network, producing question feature vectors. The model uses a question-driven attention mechanism that generates attention scores based on question relevance, focusing on essential image regions. These scores are calculated and normalized to highlight areas pertinent to the question, suppressing irrelevant visual data. A co-attentive network follows, consisting of Self-Attention (SA) and Guided Attention (GA) units. SA enhances dependencies within the question to capture critical words, while GA aligns these question features with image features, reinforcing cross-modal relevance. The attended features pass through a Multi-Layer Perceptron (MLP) with softmax to predict answers based on the most likely match. The DQMA model is evaluated on the VQA v2.0 dataset. DQMA achieves notable improvements over prior models, particularly on complex, multi-object scenes. DQMA's targeted attention mechanisms allow it to focus on relevant areas, reducing noise from extraneous data and improving answer accuracy. It outperforms other methods in capturing complex relationships in crowded scenes.

[11] The paper presents VLC-BERT, a Visual-Language-Commonsense BERT model for Visual Question Answering (VQA) that incorporates commonsense knowledge to enhance VQA performance on tasks requiring contextual reasoning. VLC-BERT targets VQA tasks that require external commonsense knowledge beyond visual understanding, aiming to answer complex questions by inferring contextual information relevant to the scene depicted in an image. The model uses Faster R-CNN for detecting and encoding image regions, and COMET (a transformer model trained on commonsense knowledge bases like ConceptNet and ATOMIC) to generate commonsense inferences related to the question. This approach allows VLC-BERT to incorporate dynamically generated contextual knowledge instead of relying on static knowledge bases. Because COMET generates numerous inferences, VLC-BERT filters them using SBERT for semantic ranking, selecting the top inferences most relevant to the question. The filtered inferences are fused with the visual and textual features using a multi-head attention mechanism that prioritizes the most helpful inferences. A single-stream transformer (VL-BERT) is then used to fuse the image regions,

question, and commonsense inferences into a unified representation. An attention-driven fusion mechanism ensures the model focuses on the most contextually relevant knowledge. VLC-BERT was evaluated on the OK-VQA and A-OKVQA datasets, which emphasize questions requiring external knowledge. VLC-BERT outperformed models that use static knowledge bases by utilizing contextualized commonsense knowledge, demonstrating superior accuracy in tasks involving complex reasoning beyond visual perception. VLC-BERT's use of contextualized commonsense inferences allows it to perform well on knowledge-intensive tasks, outperforming other models that rely on static knowledge bases. This approach is efficient in adapting to the dynamic context of the question. The reliance on COMET limits the model when questions do not require commonsense, as unnecessary inferences may introduce noise. Additionally, VLC-BERT is resource-intensive and cannot match the performance of larger models like GPT-3 in some scenarios.

[12] The UNITER model in this paper proposes a universal joint image-text representation model that can be applied to various vision-and-language (V+L) tasks, including VQA. UNITER aims to learn a universal representation for V+L tasks by pre-training on large-scale datasets with multiple objectives to create a model that can generalize across diverse tasks, including VQA, image-text retrieval, and visual entailment. UNITER uses Faster R-CNN to extract image region features and positional information. Text is tokenized and embedded using a BERT-based model, creating joint embeddings through a Transformer architecture. In Pre-training Masked Language Modeling (MLM) and Masked Region Modeling (MRM), where masking is applied to only one modality at a time, avoiding simultaneous masking. Image-Text Matching (ITM), which learns instance-level alignment between entire images and sentences. Word-Region Alignment (WRA), where Optimal Transport (OT) is used to create fine-grained alignment between words and regions, minimizing the embedding transport cost to improve alignment. Training Strategy UNITER combines these tasks across multiple datasets (COCO, Visual Genome, Conceptual Captions, and SBU Captions), integrating both in-domain and out-of-domain data to maximize generalizability. UNITER was tested on VQA, NLVR2, Visual Entailment, and Image-Text Retrieval tasks, achieving state-of-the-art results across most benchmarks. While LXMERT and UNITER both use transformer-based joint embeddings, LXMERT employs a modular two-stream architecture separating image and text processing before fusion, while UNITER uses a single-stream architecture. This distinction makes LXMERT potentially more flexible for VQA tasks, as it can handle image and text features independently. UNITER's Optimal Transport alignment, however, may enhance fine-grained image-text relationships beneficial for detailed reasoning tasks in VQA.

3. METHODOLOGY

System Architecture

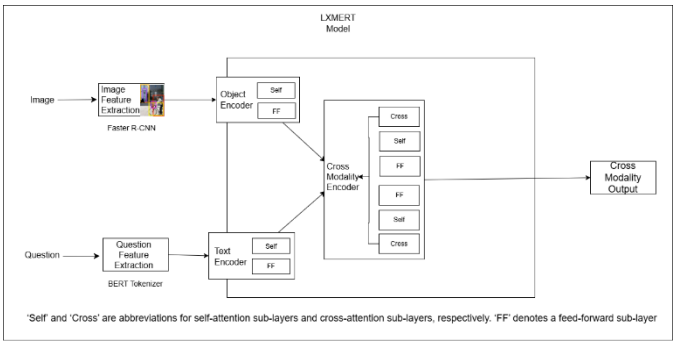

Fig - 1: Architecture Diagram

The main components of the proposed system are as follows:

- Image Feature Extraction: The model uses a pretrained object detector (e.g., Faster R-CNN) to detect objects within the image and extract feature embeddings for each object. These embeddings capture spatial, contextual, and appearance information, which are essential for understanding the image's contents.

- Question Feature Extraction: The question (text input) is tokenized, and each token is converted into embeddings using a transformer-based language model (BERT-like). This step provides contextualized word representations that capture the meaning and structure of the question.

- Object Encoder: The object encoder processes the extracted image features (object embeddings) to model relationships between detected objects. It refines spatial and relational information, enabling the model to understand interactions among objects in the image.

- Text Encoder: The text encoder processes the tokenized question to produce a contextualized representation of each word. This allows the model to understand language dependencies, such as the subject, attributes, and target objects described in the question.

- Cross Modality Encoder: This encoder applies cross-attention between the visual and text features, aligning specific words in the question with relevant objects in the image. It integrates information across modalities, linking the textual question with the visual content.

- Fusion (Fused Representation of Image & Text): The cross-modality encoder outputs a fused representation that combines both image and text features. This fused embedding captures a comprehensive understanding of the question-image pair, ready for final interpretation.

- Output Layer (Classification or Regression Layer): The fused representation is passed to the output

layer, typically a dense layer that performs classification or regression. For classification (e.g., VQA), it maps the representation to possible answers, outputting a probability distribution.

- Final Prediction: The model selects the answer with the highest probability (for classification) or outputs a continuous value (for regression), based on the question and image. This answer represents the model's best prediction, given its learned understanding from training.

4. RESEARCH GAPS AND CHALLENGES

- Dataset Limitations:
  While datasets like VQA v2.0 and Visual Genome provide extensive annotations, they still have limitations in diversity and complexity. Many existing datasets lack sufficient diversity in visual scenes and question types, which can lead to overfitting and reduced generalization in real-world applications.

- Model Generalizability:
  Transformer-based models, despite their impressive performance, often struggle with generalizing to unseen data or adapting to different domains. Ensuring that VQA models maintain high accuracy across diverse scenarios remains a significant challenge.

- Computational Constraints:
  Training large transformer models requires substantial computational resources, which can be a barrier for researchers with limited access to high-performance hardware. Additionally, deploying these models in resource-constrained environments poses challenges related to efficiency and latency.

- Bias and Fairness:
  VQA models can inadvertently learn and perpetuate biases present in training data, leading to unfair or inaccurate answers. Addressing biases and ensuring fairness in VQA systems is crucial for their ethical deployment.

5. FUTURE SCOPE

A promising future direction for this Visual Question Answering (VQA) system involves scaling it into a robust, real-time, multi-modal AI service capable of handling complex, domain-specific queries across diverse image datasets. This would mean extending the model's training beyond VQA v2.0 and Visual Genome to encompass specialized datasets from fields like medical imaging, autonomous driving, and satellite imagery. By training on vast, heterogeneous datasets, the model would be better equipped to understand and reason about complex visual content in various professional and industrial contexts.

To support this large-scale deployment, integration with powerful cloud-based platforms, such as distributed GPU clusters or TPUs offered by providers like Google Cloud and Amazon Web Services, would be essential to manage the significant computational requirements. This expanded system could then be deployed as an API, serving industries that demand advanced visual understanding capabilities, including healthcare, defense, and smart city infrastructures. The aim is to create a scalable service that can interpret and respond to specialized visual information in real time, paving the way for transformative applications across sectors.

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 atention(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.