Transformer-Driven Visual Question Answering on Complex Datasets

Stage I - Dissertation Report

Submitted by

Saurabh Madake 712352023

in partial fulfillment for the award of the degree

of

M. Tech. (Computer Engineering)/ (Data Science)

Under the guidance of

Prof. Mrs. Archana Patil

COEP Tech, Pune

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

COEP TECHNOLOGICAL UNIVERSITY (COEP Tech)

PUNE

05 October 2024

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

COEP TECHNOLOGICAL UNIVERSITY (COEP Tech) PUNE

CERTIFICATE

Certified that the Stage - I dissertation titled, "Transformer-Driven Visual Question Answering on Complex Datasets" has been successfully completed by

Saurabh Madake 712352023

and is approved for the partial fulfillment of the requirements for the degree of "M.Tech. Computer Engineering - Data Science".

SIGNATURE	SIGNATURE
Prof. Mrs. Archana Patil	**Dr. P.K. Deshmukh**
Project Guide	**Head**
Dept. of Computer Science	**Dept. of Computer Science**
& Engineering	**& Engineering**
COEP Technological University,	**COEP Technological Univesity,**
Shivajinagar, Pune - 5.	**Shivajinagar, Pune - 5.**

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.

Contents

Chapter 1

Introduction

Visual Question Answering (VQA) is a field that bridges computer vision and natural language processing (NLP) to enable machines to understand and respond to questions about visual content. Hence it is also known as an AI Complete task. The significance of VQA lies in its potential applications, such as assisting visually impaired individuals, enhancing human computer interactions, etc. with this project I am aiming to explore the evolution, current state, and future directions of VQA, with a focus on transformer based model like LXMERT, the utilization of datasets such as VQA v2.0 and Visual Genome, and the implementation frameworks like PyTorch and Hugging Face. Unlike traditional methods which required a separate pipeline for textual and visual data, recent development in transformers like LXMERT have revolutionize this process. Transformers excel at integrating multimodal data. this helps in complex tasks like VQA.

Chapter 2

Literature Review

Key Studies

Sr. No	**Author Name**	**Paper Title**	**Date of publishing**	**Observations**

1	Tan and Bansal	LXMERT: Learning Cross-Modality Encoder Representations from Transformers	2019	Introduced the LXMERT model, showing significant improvements in VQA tasks by integrating visual and textual information. LXMERT uses three types of encoders: object-relationship, language, and cross-modality. Evaluated on datasets like VQA, GQA, and NLVR2, it outperformed prior models, showing superior performance in cross-modality tasks compared to previous BERT-like models focused solely on language.

2	Liunian Harold Li et.al	VisualBERT: A Simple Model for Vision-and-Language Tasks	2019	VisualBERT integrates BERT with image features extracted from object detectors, using Transformers to process text and image inputs. It undergoes task-agnostic pre-training, task-specific pre-training, and fine-tuning for specific tasks. Pre-trained on COCO, it serves as an effective baseline for vision-language tasks, leveraging BERT's language modeling capabilities.

3	Lu et al.	ViLBERT: Pretraining Task-Agnostic Vision-and-Language BERT	2019	ViLBERT extends BERT with a two-stream architecture that processes textual and visual inputs separately and then interacts through co-attentional transformers. Pretrained on Conceptual Captions dataset, it is optimized for vision-language tasks like visual question answering and visual commonsense reasoning, outperforming task-specific models by achieving state-of-the-art results.

4	Kaavya Rekanar et. al	Optimizing Visual Question Answering Models for Driving: Bridging the Gap Between Human and Machine Attention Patterns	2024	This paper focuses on the attentional gap between humans and machines in driving scenarios for VQA models. The proposed filter enhances the attention mechanism in LXMERT for driving-specific elements, such as roads and vehicles. The model's performance improves significantly in driving scenarios by aligning attention more closely with human observation patterns.

5	Zhicheng Huang et. al	Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers	2020	Pixel-BERT introduces a deep multi-modal transformer architecture to improve alignment between text and image pixels. By using image-sentence pairs, it improves the semantic links between language and visual data beyond region-based features. It achieves state-of-the-art results across several tasks like VQA, showing a 2.17-point improvement compared to previous models.

6	Wang, Jianfeng et. al	Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering	2024	Med-VQA combines large language models and multi-label learning to address data scarcity and complexity in Medical VQA. The Image to Label to Answer (ITLTA) framework reduces costs, improves interpretability, and enables zero-shot learning. Experiments on the VQA-Med 2019 dataset show that Med-VQA outperforms current approaches, enabling more effective clinical applications.

7	Rub'en Tito et. al	Hierarchical multi-modal transformers for Multi-Page DocVQA	2023	This paper extends Document VQA to multi-page documents with the MP-DocVQA dataset and proposes Hi-VT5, a hierarchical multimodal transformer. Hi-VT5 processes multi-page documents effectively and outperforms existing models in terms of performance and explainability, making it a better fit for real-world applications involving multi-page document contexts.

8	Aisha Urooj Khan et. al	MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering	2020	In order to tackle the problem of Visual Question Answering (VQA), the MMFT-BERT model (Multimodal Fusion Transformer with BERT encodings) processes various modalities—text, video, and subtitles—individually and collaboratively. Each modality is given its own BERT encoder, and the outputs are then fused together using a brand-new transformer-based fusion method. On the TVQA dataset, the model performs better than earlier state-of-the-art models, demonstrating gains in accuracy.

9	Wonjae Kim et.al	ViLT: Vision and Language Transformer Without Convolution or Region Supervision	2021	A vision-language pre-training paradigm called ViLT (Vision and Language Transformer) removes the requirement for region supervision and convolutional neural networks (CNNs) to streamline visual processing. ViLT embeds picture patches using a linear projection, in contrast to typical models that rely on object detectors and CNNs for image feature extraction. This architecture maintains competitive performance across vision-language tasks such as VQAv2 and image retrieval, while reducing computational complexity to make it up to ten times faster than previous models.

10	Tim Siebert et. al	Multi-Modal Fusion Transformer for Visual Question Answering in Remote Sensing	2022	The paper presents a new architecture for visual question answering (VQA) in remote sensing (RS) called VBFusion, which jointly learns image and text representations by using multimodal transformer models. By utilizing a feature extraction module, a fusion module that combines multiple VisualBERT layers, and a classification module, this approach overcomes the limitations of existing models and can handle complex, non-specific questions that go beyond predefined object categories. The results of the experiment show notable gains in performance on RS VQA datasets.

11	Jinmeng Wu et. al	Question-Driven Multiple Attention (DQMA) Model for Visual Question Answer	2022	The Question-Driven Multiple Attention (DQMA) model is proposed in this paper. It deals with the problem of extraneous visual information influencing VQA models' accuracy. The DQMA model uses LSTM for question features and Faster R-CNN for image feature extraction. Relevant image regions are chosen by a question-driven attention mechanism, which lowers noise. In order to improve the interaction between the question and image features, the model also incorporates co-attention networks. Assessments conducted on the VQA 2.0 dataset reveal that the DQMA model surpasses alternative techniques, enhancing overall precision.

12	Zekai Shao et. al	Visual Explanation for Open-Domain Question Answering With BERT	2024	In order to clarify the decision-making process in Open-Domain Question Answering (OpenQA) models, especially those that use BERT, the paper suggests VEQA, a visual analytics system. OpenQA addresses issues with complicated data and models while providing answers to queries derived from lengthy unstructured text passages. In order to address these, VEQA provides visual explanations at three different levels: summary, instance, and module. This enables experts to examine the way in which models handle questions and extract pertinent passages. By using visual aids such as ranking visualizations and comparative trees, VEQA enhances the interpretability of models and identifies areas that need improvement.

13	Sahithya Ravi et. al	VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge	2023	The VLC-BERT model is designed for Visual Question Answering (VQA) with an emphasis on commonsense reasoning. It incorporates contextualized knowledge using the COMET model, which is trained on human-curated knowledge bases like ConceptNet and ATOMIC. The model combines visual and textual inputs with commonsense inferences to answer questions that require reasoning beyond what is visible in the image. Through attention mechanisms, VLC-BERT selects the most relevant commonsense knowledge to enhance the VQA task.

14	Dhiraj Amin et. al	Visual Question Answering System for Indian Regional Languages	2022	The paper "Visual Question Answering System for Indian Regional Languages" addresses the scarcity of datasets for Indian languages like Hindi and Marathi in visual question answering (VQA) systems. The authors investigate the adaptation of the easy-VQA dataset for these languages, developing models that integrate natural language processing and computer vision to answer questions based on images. Various architectures, including CNNs and RNNs like LSTM, are explored for handling both image and text inputs. The paper demonstrates how translated datasets and deep learning techniques can bridge the gap for regional language VQA tasks.

15	Venkat Kodali et. al	Recent, Rapid Advancement in Visual Question Answering: a Review	2022	The paper ”Recent, Rapid Advancement in Visual Question Answering: a Review” provides a detailed overview of the rapid developments in Visual Question Answering (VQA). It highlights how VQA, combining computer vision and natural language processing, has grown exponentially in research output since 2015. The review covers key VQA models and techniques, including attention-based and transformer architectures, particularly BERT, which have revolutionized VQA performance. It also discusses medical image datasets and their unique VQA challenges, such as data scarcity, and offers suggestions for future research directions.

2.1 Current Trends

2.1.1 Multimodal Transformers

Recent advancements emphasize the development of multimodal transformers that can seamlessly integrate and process information from both visual and textual modalities. Models like UNITER and ViLT have pushed the boundaries of VQA by enhancing cross-modal understanding and reasoning capabilities.

2.1.2 Attention Mechanisms

Attention mechanisms remain a cornerstone in VQA models, enabling them to focus on relevant parts of the image based on the question. Innovations in attention, such as self-attention and cross-attention layers, have improved the interpretability and accuracy of VQA systems.

2.1.3 Data Augmentation and Synthetic Data

To address data scarcity and enhance model robustness, researchers are increasingly using data augmentation techniques and synthetic data generation. Approaches like GAN-based image synthesis and question generation have been employed to expand existing datasets and introduce variability, thereby improving model generalization.

2.2 Gaps and Challenges

2.2.1 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.

2.2.2 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.

2.2.3 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.

2.2.4 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.

Chapter 3

Motivation

In today's world Artificial intelligence is developing at a rapid pace, which has increased demand for systems that can interact and understand visual and textual information. The challenge of Visual Question Answering (VQA) for models that can integrate natural language processing and vision, potentially resulting in more reliable and understandable artificial intelligence systems. Applications of VQA are found in many different fields, such as robotics, autonomous vehicles, assistance for the blind, and educational resources. The potential for AI to effectively interpret and react to language and visual queries can significantly improve human-computer interaction therefore, VQA is an AI-complete task with great potential for future technological advancement. The study intends to advance this field by investigating the efficacy of sophisticated models, like LXMERT, in solving VQA challenges, ultimately leading to more efficient and reliable AI solutions.

Chapter 4

Problem Statement

Transformer-Driven Visual Question Answering on Complex Datasets

Visual Question Answering (VQA) poses a significant challenge in artificial intelligence due to the need for seamless integration of visual and textual data. Traditional models often struggle to establish connections between image content and associated queries, leading to suboptimal performance. Existing approaches also face difficulties in understanding complex image regions and providing accurate answers to diverse and nuanced questions. To address these challenges, recent advancements in transformer-based architectures, such as LXMERT, offer new possibilities by leveraging deep contextual understanding across modalities.

Despite their potential, there remains a need to evaluate their performance on large-scale datasets, such as VQA v2.0 and Visual Genome, to understand their limitations and capabilities in handling real-world questions.

This dissertation aims to develop a transformer-based VQA system that improves the interpretability and accuracy of responses by fusing image and textual representations more effectively. The goal is to assess whether transformers can bridge the gap in current VQA approaches and enhance overall system performance across complex queries.

Chapter 5

Approach

This project focuses on using the LXMERT model, a transformer architecture specifically designed for multi-modal tasks, to address the challenges in Visual Question Answering (VQA). LXMERT excels at learning the alignment between visual data and text, making it well-suited for this task. The project integrates two large datasets—VQA v2.0 and Visual Genome—to enrich the visual and textual information available for model training.

5.1 SDLC Phases

1. **Requirement Gathering:**
    - The project aims to build a VQA system that leverages the LXMERT model to fuse visual and textual features for generating accurate responses to image-based questions. The datasets include VQA v2.0 for image-question-answer pairs and Visual Genome for additional object relationships and attributes.
2. **Dataset Combination:**
    - The images and question-answer pairs between VQA v2.0 and Visual Genome are combined and used to enrich the dataset. Also, objects and relationships

from Visual Genome are used to train the model. This combination helps improve the model's performance, particularly on more complex, relational queries.

3. **Implementation:**

   - **Data Preprocessing:** Questions are tokenized, and object attributes and relationships from Visual Genome are appended to the corresponding image-question pairs. This creates a combined dataset that provides richer information for the LXMERT model.

   - **Model Training:** The LXMERT model is fine-tuned using the combined dataset. Visual features and text embeddings are processed through LXMERT's cross-attention layers, which align the visual and textual modalities. This model is expected to handle both simple and complex queries by understanding the relational context provided by the Visual Genome dataset.

4. **Testing:**

   - The model is evaluated using accuracy and other metrics like precision, recall, and F1-score on a test split of the VQA v2.0 dataset. The effectiveness of the dataset combination is measured by comparing the model's performance on the original VQA v2.0 dataset.

5. **Deployment:**

   - The final trained model will be deployed locally using a web interface. This UI allows users to upload images and ask a question about the image, with the LXMERT model providing answers based on the visual and textual inputs.

5.2 Dataset Combination and Utilization:

1. **VQA v2.0 Dataset:**

    - VQA v2.0 provides image-question-answer triplets, which test the model's ability to understand and reason about the contents of an image. This dataset is the foundation for building the VQA system.

2. **Visual Genome Dataset:**

    - Visual Genome offers scene graphs, which capture relationships between objects, attributes, and interactions within images. These relationships help in answering more nuanced questions that require relational understanding or contextual details.

3. **Combining the Datasets:**

    - Images between VQA v2.0 and Visual Genome, the question-answer pairs from VQA v2.0 are enhanced with additional context from the Visual Genome dataset. This includes relationships between objects, their attributes, and spatial configurations, which help the model better understand complex scenarios.

4. **Utilization:**

    - The LXMERT model processes both visual features and text embeddings using its cross-attention layers to create a fused representation of the image and the question. The combined dataset, enriched by Visual Genome's relational data, allows the model to improve its understanding and provide more accurate responses, particularly for complex questions involving spatial or relational reasoning.

Chapter 6

Research Objectives

1. To develop a Visual Question Answering system utilizing the LXMERT model, which effectively integrates visual and textual data for answering image-based questions.

    - Focusing on building a system that efficiently processes both visual content from images and textual content from questions, using LXMERT's cross-attention mechanism to align the two modalities.

2. To evaluate the performance of the LXMERT model on the VQA v2.0 dataset and compare it with other baseline models in terms of accuracy and efficiency.

    - This involves benchmarking LXMERT against traditional VQA models, assessing its ability to answer a wide range of questions across diverse image types, and measuring its overall effectiveness.

3. To investigate the impact of combining VQA v2.0 with Visual Genome data, specifically focusing on how enriched object relationships and attributes from Visual Genome improve question-answering performance.

    - Exploring how the additional contextual and relational information from Visual Genome enhances the model's ability to tackle more complex queries that require understanding object interactions and spatial relationships.

4. To enhance the capabilities of visual question answering (VQA) systems

    - Improving their accuracy and efficiency through the exploration of advanced multimodal transformer architectures, while providing insights and guidelines for handling multimodal data that can inform future research in VQA and related fields.

Chapter 7

Hardware and Software Requirements

7.1 Hardware Requirements

- **Primary GPU**: NVIDIA GeForce GTX 1650 (for model training and inference).
  - CUDA cores: 1024
  - VRAM: 4 GB
- **Secondary GPU**: Intel Iris Xe (for lighter tasks and preprocessing).
- **Processor**: Intel Core i7 for handling large dataset preprocessing tasks.
- **RAM**: 16 GB for handling large datasets and model training.
- **Storage**: 500 GB SSD for datasets, checkpoints, and logs.

7.2 Software Requirements

- **Operating System**: Windows 10 or Linux
- **Programming Language**: Python 3.8+

- **Development Environment**: Visual Studio Code
- **Libraries/Frameworks**:
  - **PyTorch**: PyTorch has become a preferred framework for implementing deep learning models due to its dynamic computation graph and ease of use.
  - **Hugging Face**: Hugging Face's Transformers library offers pre-trained models and tools that streamline the development of transformer-based architectures.
  - **Transformer**: LXMERT (Learning Cross-Modality Encoder Representations from Transformers) is a VQA model with separate language and vision encoders, followed by cross-modality encoders to capture complex relationships between visual elements and textual queries, enhancing VQA performance.
  - **CUDA**: For leveraging NVIDIA GPU during training.
  - **Flask/Streamlit**: For building the user interface for image and question input.

7.3 Datasets

7.3.1 VQA v2.0 Dataset

- **Number of Images**: 204,721
  - Training Set: 82,783 images
  - Validation Set: 40,504 images
  - Test Set: 81,434 images
- **Number of QA Pairs**: 1,105,904
  - Training Set: 443,757 questions
  - Validation Set: 214,354 questions

- Test Set: 447,793 questions

- **Questions per Image**: 3 questions per image (on average)
- **Answers per Question**: 10 answers per question (annotated by 10 human annotators)
- **Usage**: Provides the primary image-question-answer pairs for training and testing the model.
- **Format**: JSON (for question-answer pairs) and image files.

7.3.2 Visual Genome Dataset

- **Number of Images**: 108,249
- **Number of QA Pairs**: 1.7 million question-answer pairs
- **Questions per Image**: About 17 questions per image
- **Types of QA Pairs**:
  - Region-based QA (localized to regions in the image)
  - Attribute QA
  - Relationship QA
  - Scene Graphs: Available for all images, with objects, relationships, and attributes annotated for each image.
- **Usage**: Enhances the VQA v2.0 dataset by providing additional object, relationship, and attribute information.
- **Format**: Scene graphs, relationships, attributes (JSON) and image files.

Chapter 8

Proposed System Design Components

Image Preprocessing and Feature Extraction

Purpose: Extract object-level features from input images to feed into the LXMERT Object Encoder.

Steps:

1. Load images from the dataset (e.g., VQA v2.0 or Visual Genome).
2. Preprocess images (resize, normalize, convert to tensors).
3. Pass preprocessed images through a pre-trained Faster R-CNN model.
4. Extract features such as bounding boxes and region embeddings.
5. Save extracted features to `.pkl` files for faster access during training.

Question Preprocessing and Feature Extraction

Purpose: Convert natural language questions into tokenized representations for the Text Encoder.

Steps:

1. Tokenize questions using the LXMERT tokenizer (provided by Hugging Face).
2. Generate positional embeddings for question tokens.
3. Feed tokenized data to the LXMERT Text Encoder.

Object Encoder

Purpose: Encodes image features into latent representations suitable for cross-modal interactions.

Steps:

1. Receive image features (bounding boxes, region embeddings) as input.
2. Pass them through attention layers to encode object-level context.

Text Encoder

Purpose: Encodes tokenized questions into contextual representations.

Steps:

1. Input tokenized questions and positional embeddings.
2. Use transformer-based attention layers to capture linguistic context.

Cross-Modality Encoder

Purpose: Perform cross-modal attention between encoded image and question features.

Steps:

1. Use self-attention and cross-attention layers to combine object and text encodings.

2. Output joint representations of image and text modalities.

Prediction Head

Purpose: Predict the answer to the input question based on the cross-modal representation.

Steps:

1. Receive joint representations from the cross-modality encoder.

2. Pass through a classifier or fully connected layer.

3. Generate the predicted answer.

Evaluation and Fine-Tuning

Purpose: Assess model performance and adapt to different datasets.

Steps:

1. Evaluate using validation splits (e.g., VQA metrics like accuracy).

2. Fine-tune on domain-specific data if needed.

Chapter 9

Current Implementation Status

Work Completed to Date

1. **Literature Review**
    - Conducted a thorough review of research papers and existing implementations of LXMERT for VQA tasks.
    - Explored the capabilities of scene graphs, object relationships, and attributes in Visual Genome to enrich question-answer pairs.
    - Created a detailed survey paper that summarizes and analyzes existing research on the topic.
2. **Environment Setup**
    - Configured a development environment using Visual Studio Code with CUDA support to accelerate model training on an NVIDIA GTX 1650 GPU.
    - Resolved various configuration issues related to PowerShell script execution.
3. **Dataset Preparation**
    - Downloaded and analyzed the VQA v2.0 dataset and Visual Genome datasets.

- Preprocessed the image features using a Faster R-CNN model, saved in a `.pkl` file to overcome memory and computational limitations.

Challenges Faced

- **Hardware Constraints**: Encountered memory crashes in Visual Studio Code due to the size of the datasets during loading.
- **Model Complexity**: LXMERT's architecture required fine-tuning to ensure compatibility with preprocessed inputs.
- **Dataset Issues**: Ensuring alignment of visual and textual modalities during preprocessing required additional debugging.

Decisions Made

- Opted for separate training loops for VQA v2.0 and Visual Genome datasets to better manage resource constraints and dataset complexity.
- Prioritized preprocessing to streamline training on limited hardware resources.

Work to be Done

1. **Implementation of the Model Pipeline**
    - Integrate the Hugging Face transformers library and PyTorch framework to build and evaluate the LXMERT model.
2. **Model Training**
    - Begin training the LXMERT model on the preprocessed VQA v2.0 data.

- Monitor training metrics such as loss and accuracy to identify any convergence issues.

3. **Validation and Testing**
   - Validate the trained model on a reserved portion of the dataset.
   - Test the model with real-world question-image pairs to evaluate its robustness and performance.

4. **Integration with Visual Genome Dataset**
   - Preprocess and align the Visual Genome dataset using the existing pipeline.
   - Train the model on Visual Genome data, leveraging the insights from VQA v2.0 training.

5. **Optimization**
   - Experiment with hyperparameter tuning to improve performance.
   - Optimize data loading and training to mitigate memory consumption issues.

6. **UI Development**
   - Design a simple web-based interface for inputting images and questions, displaying answers generated by the model.

7. **Documentation and Reporting**
   - Continue maintaining detailed documentation of challenges, decisions, and progress.

Chapter 10

Future Scope

A concrete future plan for this Visual Question Answering (VQA) system is to scale it into a large-scale, real-time, multi-modal AI service capable of handling complex, domain-specific queries across a wide variety of image datasets. This would involve expanding the model's training on vast, heterogeneous datasets far beyond VQA v2.0 and Visual Genome, incorporating specialized datasets from diverse fields such as medical imaging, autonomous driving, and satellite imagery.

Additionally, this future iteration would require integration with large-scale cloud-based platforms, such as distributed GPU clusters or TPUs provided by companies like Google Cloud or Amazon Web Services, to manage the immense computational demand.

The aim would be to build a service that could be deployed as an API, serving industries that require advanced visual understanding and reasoning capabilities, such as healthcare, defense, and smart city infrastructures.

Chapter 11

References

[**1**] Tan, H. and Bansal, M., 2019. LXMERT: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.

[**2**] 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.

[**3**] 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.

[**4**] 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.

[**5**] 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.

[**6**] 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.

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

[**8**] 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.

[**9**] 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.

[**10**] 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.

[**11**] 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.

[**12**] Shao, Z., et al., 2024. Visual explanation for open-domain question answering with BERT. IEEE Transactions on Visualization and Computer Graphics, 30(7), pp. 3779-3797. https://doi.org/10.1109/TVCG.2023.3243676.

[**13**] 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.

[**14**] Amin, D., Govilkar, S., and Kulkarni, S., 2022. Visual question answering system for Indian regional languages. 2022 5th International Conference on Advances in

Science and Technology (ICAST), Mumbai, India, pp. 22-27.

[**15**] Kodali, V. and Berleant, D., 2022, May. Recent, rapid advancement in visual question answering: A review. In 2022 IEEE International Conference on Electro Information Technology (EIT) (pp. 139-146). IEEE.

[**16**] 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.

[**17**] 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.