A Project on

Generative AI for Question Answering

*Submitted in partial fulfilment of the requirement for the award of the degree of*

Master Of Computer Application

A blue and red text on a white background

Description automatically generated

Under The Supervision of

Dr. Arun Chaudhary

Submitted By

Deepak Garg

Diksha Bisht

Nishka Mehlawat

SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY

BENNETT UNIVERSITY, GREATER NOIDA, U.P

# A blue and red text on a white background Description automatically generatedSchool of Computer Science Engineering and Technology - Bennett University

CANDIDATE’S DECLARATION

I/We hereby certify that the work which is being presented in the project, entitled “ Generative AI for Question Answering” in partial fulfilment of the requirements for the award of the Maters of Computer Application submitted in the School of Computing Science and Engineering of Bennett University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Dr. Arun Chaudhary, School of Computer Science Engineering and Technology, Bennett University, Greater Noida

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

Team Member Name

Deepak Garg - e23mcag0037

Diksha Bisht – e23mcag0049

Nishka Mehlawat-e23mcag0020

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Table of Contents

Title Page No.

|  |  |  |
| --- | --- | --- |
| Abstract |  | I |
|  |  |  |
|  |  |  |
| Chapter 1 | Introduction   * 1. Motivation   2. Problem statement   3. Aim and objectives of the study. | 1 |
| Chapter 2 | Literature Survey   * 1. A review of relevant literature and existing work related to the project. | 2 |
| Chapter 3 | Project Design  Hardware/ Software Requirement   * 1. Proposed Methodology   2. Details on data collection methods and tools   3. Procedures and techniques used in the project | 4 |
| Chapter 5 | Conclusion and Future Work  References | 17 |

Abstract

A promising method for question answering (QA) and visual question answering (VQA) is generative AI. Generative models seek to produce the answer directly from the input question and context, in contrast to classic discriminative models, which concentrate on choosing the right response from a range of options. Generative models can be trained to produce logical and fluid responses to open-ended inquiries in the context of quality assurance. Similar to this, generative models can be trained in VQA to produce responses that precisely characterise an image's content and address associated queries.

Large amounts of high-quality training data are necessary, and this is one of the main issues in developing generative models for QA and VQA. In order to overcome this difficulty, scientists have created a number of data augmentation methods, including back-translation and paraphrase, to produce more training instances. The requirement for efficient assessment measures that can precisely gauge the calibre of produced responses is another difficulty. In order to achieve this, scientists have put forth a number of metrics, including BLEU, ROUGE, and METEOR, which contrast the produced responses with human-provided reference replies.

Keywords:- Question Answering, Generative AI, BLEU, ROUGE, METEOR

1. INTRODUCTION

In Natural Language Processing (NLP) and Computer Vision (CV), respectively, Question Answering (QA) and Visual Question Answering (VQA) are two major study fields. Although VQA seeks to respond to inquiries about visual content, such as photos, QA's main objective is to create systems that can automatically respond to queries in natural language.

Information retrieval (IR) methods are commonly employed by QA systems to locate pertinent texts or sections that might hold the solution to a particular query. These systems then take the detected text and extract the answer using natural language processing (NLP) techniques including dependency parsing, Named Entity Recognition (NER), and Part-of-Speech (POS) tagging. Recurrent Neural Networks (RNNs) and Transformers are two deep learning-based techniques that have been utilised to create more complicated quality assurance (QA) systems that can manage complex queries and context-dependent responses.

Numerous real-world uses of QA and VQA exist, such as chatbots, virtual assistants, customer support, and education. Nonetheless, creating precise and dependable QA and VQA systems continues to be a difficult research topic because these systems need to be able to manage a variety of question kinds, languages, and domains in addition to dealing with problems like bias, ambiguity, and hostile attacks.

1.1 INTENDED AUDIENCE

Developers, researchers, and data scientists interested in creating and experimenting with generative AI models for question answering and visual question answering are the target audience for this project. This project requires a basic familiarity of natural language processing methods, deep learning frameworks like TensorFlow and PyTorch, and machine learning ideas. It is also intended for people who have worked with sizable datasets before and have knowledge of model training and data pre-processing. Professionals in the business that want to use generative AI for chatbots, virtual assistants, and customer service may also find this project interesting.

1.2 PROJECT SCOPE

The goal of this project is to create a generative AI model for answering questions and providing visual answers. Researchers, developers, and data scientists interested in learning more about the potential of generative AI models for natural language processing applications are the target audience for this project. The project will include a number of technical components, such as evaluation, model training, and data pre-processing.

Creating a generative AI model that can produce answers to queries based on textual and visual inputs is the goal of the model training phase. To do this, the study will investigate a variety of designs, including transformer-based models. A sizable dataset of question-answer pairs and visual data will be used to train the model, and the project will look into methods for optimising the model's performance.

This evaluation will include both quantitative analysis using pre-established criteria and qualitative examination of the produced responses. Ethical issues, such as possible biases and privacy problems, will be closely monitored during the research.

1. LITERATURE SURVEY

In previous research, Table 1. Summarization of literature review describe different part of components and techniques used by various researchers. The field of question answering (QA) provides important insights and approaches that act as stimulants for more research in this area. These include dataset creation, model architecture design, attention mechanisms, integration of external knowledge, and performance evaluation, by addressing important issues and offering creative solutions, these studies greatly advance QA .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author | Techniques | Dataset | Methods | Accuracy |
| [1], [2] | CBM , OSCAR , MMBERT | OK-VQA [10], VQA 2.0 | Text- only Approach | 67.8% , 67.7% |
| [3] | BLEU , VIT | VQA-RAD , Path VAQ | Multi-model fusion | 83.86% , 62.37% |
| [4] | IQAN , MUTAN | CLEVR , VQA 2 | Dual training schema | 0.88% |
| [5] | MCB-A , MLP | COCO-VQA , TDIUC | Modular Approach | 86.3% , 95.02% |
| [6] | BLIVA | YouTube thumbnails | VLM Approach | 83.5% |
| [7] | MFH , MFB | OK-VQA , KVQA , FVQA | Knowledge based Approach | 59% , 69% |
| [8], [9] | MCAN , RAMEN | CLEVR , VQA-V2 | Joint embedding , Attention-based , Compositional model , External-Knowledge based. | 98.4% , 70.63% |
| [10] | iBOWIMG | DAQUAR , COCO-VQA | Non-Deep Learning , Attention based | 29.30% , 68.60% |
| [11], [12] | LSTMQ , MLP | MC COCO, MS COCO | Baseline method , join parsing | 58.16% , 63.09% |
| [13] | LXMERT , UNITER , SOHO | MS COCO | Transformer based VQA | 70.5% |
| [14] | Dual QA | SQUAD | regularization | 70% |
| [15] | LSTM | MS MARCO | Transformer encoder decoder | 64.52% |

Table 1. Summarization of literature review

Salaberria et al. [1], [3] proposed a framework comparably, the OSCAR model adds object tags to the input and suggests various pretraining approaches. The two steps that make up our caption-based model, or CBM, are as follows: (i) a system for creating captions that provides a succinct explanation for an image and (ii) a language model that uses the caption and an inquiry to provide an answer. Bazi et al. [3] suggested the experiments, two VQA datasets for radiology images called VQA-RAD and PathVQA are used to validate the model. The model's performance is evaluated using a number of metrics, such as BLEU scores and open-ended and closed-ended accuracy ratings obtained from the use of two decoder layers. The model's predictions and the ground-truth answers may not overlap as much, according to the lower BLEU scores.

Li et al. [4] introduced a visual attention mechanism to draw attention to the areas of the image that are important for answering the question. One of the most advanced approaches to modelling interactions is the bilinear model MUTAN. Utilise the ideas in the field of computer vision and suggest CycleGAN for the unsupervised learning of image-to-image translation functions.

Kafle et al.[5] presented three sources are cited in the questions. Initially, imported a portion of the COCO-VQA and Visual Genome questions. Second,  developed algorithms to generate questions based on the semantic segmentation annotations from COCO  and the objects and attributes annotations from Visual Genome . Third, for certain question types, we employed human annotators. More capable model MCB performs worse for absurd; however, the version trained without absurd exhibits significantly smaller differences than Q+I, indicating MCB's superior ability to recognise absurd questions. The twelve question types in TDIUC were selected to reflect both traditional computer vision tasks and new, advanced vision tasks that call for various levels of image comprehension and reasoning.

Hu W et al.[6] proposed BLIVA, a multimodal LLM that combines image-encoded patch embeddings, which contain richer image information, with learned query embeddings, which are more closely aligned with the LLM. Ahir et al. [7]introduced knowledge base file to choose our model's accuracy and various parameters. The OkVQA Dataset has an average of 14,055 open-ended questions with five ground truth answers per question.

Mostafa et al. [8] suggested RAMEN model stands for Recurrent Aggregation of Multimodal Embeddings Network. The model processes the image and question features in several stages. In the first stage, concatenation is used to combine each image feature vector which are the feature vectors of region proposals with the question vector. MCAN, an alternative attention method, employs a layer known as the MCA (Modular Co Attention) layer. The self-attention unit and the guided attention unit make up the two components that make up the MCA layer, which is why the method is named modularly.

Gupta et al. [10] presented iBOWIMG as a baseline model for VQA to extract image features, they use the output of a later layer of the pre-trained GoogLe Net model for image classification. Agrawal et al. proposed a framework. He et al. [13] In this paper, it was suggested to compare Transformer-based pretraining models and use VQA tasks for fine-tuning. In particular, the study compares two state-of-the-art models, UNITER and LXMERT, to explore the possibility of transformer-based models to improve VQA performance. In order to visually token the visual context and ultimately convert to a purely semantic representation, SOHO encourages learning a VD.

Tang et al.[14] provide a training methodology that explicitly uses the probabilistic correlation between the QA and QG models to direct their training, training them both at the same time. We put into practice a QA model based on recurrent neural networks and a QG model based on sequence-to-sequence learning. Since every element in the QA and QG models is differentiable, back propagation might be used to learn all of the parameters in these two models in a traditional manner.

1. Project Design
   1. External Interface Requirements
      1. Hardware Interfaces
2. Processing power: A large amount of processing power is usually needed for generative AI models that answer questions and answer questions visually. A high-performance CPU or GPU may be part of this, depending on how it is implemented.
3. Memory: In order to store and process data, the models utilised in this project may need a lot of memory. This could apply to RAM as well as disc space.
4. Network connectivity: High-speed network connectivity can be required if the project calls for utilising cloud-based services or gaining access to sizable databases.
5. Compatibility: TensorFlow or PyTorch are two examples of software frameworks and libraries that should function with the hardware utilised in this project.
   * 1. Software Interfaces
6. Python and its libraries, including TensorFlow, Keras, PyTorch, and OpenCV, will be the main programming language used in this project's software interface.
7. Pre-trained models must be used for computer vision and natural language processing tasks in this project.
8. Other tools needed for the project include Pandas, NumPy, and Matplotlib for data pre-processing and visualisation.
9. Methodology

This comprehensive methodology Fig2. Shows methodology integrates image processing, document processing, model configuration, initialization, inference, and evaluation stages to enable multimodal question answering capabilities leveraging both visual and textual information sources.

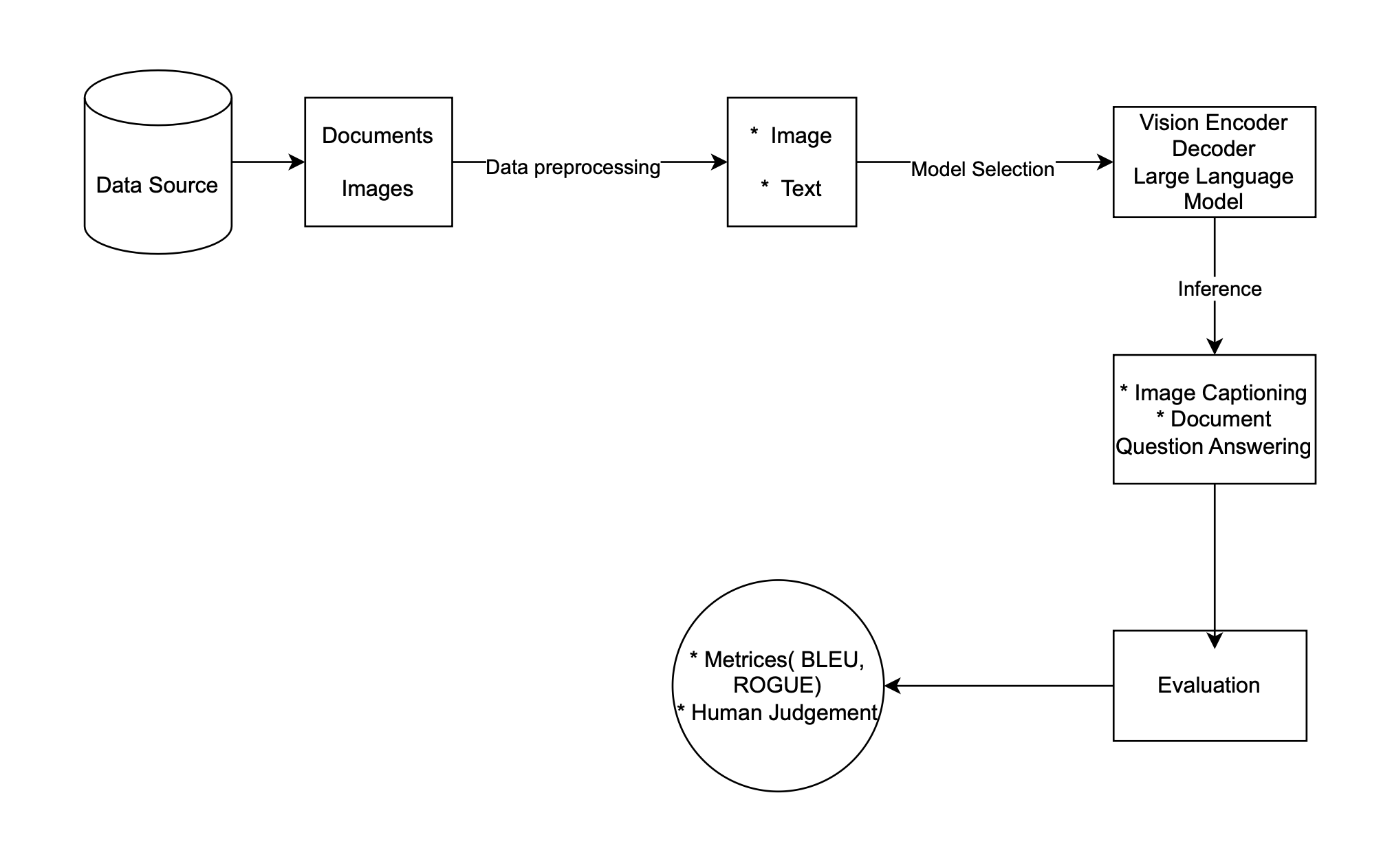


Fig 2. Methodology

Data Collection:

Several Images, URLs and documents were collected from various sources for testing the working of the model. While utilizing the model for any industry-specific work the user can enter desired data in any format and from multiple sources to process data in respective pipelines.

Image Processing:

To question answers over images, Vision Encoder-Decoder architecture is used, it combines the capabilities of Vision Encoder i.e., ViT for processing images and Bidirectional Encoder Representations from Transformers i.e., BERT model to understand language while question answering.

Document Processing:

Different documents and URLs were given as input to the model and with the help of document retrieval modules text data was collected and send for processing. I the preprocessing part unnecessary spaces and characters were removed to maintain consistency and accuracy in the data. The Large Language Model makes use of Gemini, BAAI/bge-small-en (a model form hugging face hub) framework for the question-answering and Chroma db storing and retrieval of document.

Model Configuration:

The VisionEncoderDecoderConfig was utilized to configure the vision encoder-decoder model for image captioning tasks. This configuration encompasses parameters for both the vision encoder (e.g., ViT) responsible for processing visual input and the language decoder (e.g., BERT) responsible for generating textual captions. Similarly, the configuration of the LLM for document-based question answering was established using appropriate settings.

Model Initialization:

The LLM for question answering and the vision encoder-decoder model were initialized either using pre-trained weights or from scratch. Leveraging pre-trained weights offers the advantage of transferring knowledge learned from large-scale datasets, enhancing the performance of both models in their respective tasks.

Inference:

This system utilizes two specialized models for tasks involving images and documents. For image captioning, a vision encoder-decoder model acts as an image translator, automatically generating captions based on its visual understanding. Likewise, for document-based question answering, a large language model functions as a powerful information retrieval system, searching through indexed documents to find answers to user queries.

Evaluation:

The quality of generated captions for images and answers for documents was evaluated using a combination of quantitative metrics such as BLEU score, ROUGE score, and qualitative assessments based on human judgment. These evaluations provide insights into the performance and effectiveness of the models in their respective tasks, guiding further refinement and improvement efforts.

1. Experiment

This project aims to test the Large Language Model’s capabilities to answer the questions asked using various documents and websites. It can accept URLs, PDF files, word files, .txt files and image URL as the source of domain knowledge to answer.

The project is divided into 2 major parts:

a. Question answering over text data, and

b. Question answering over Image Data

A. In the first part, the Large Language Model makes use of Gemini, BAAI/bge-small-en (a model form hugging face hub) framework for the question-answering and Chroma db storing and retrieval of document.

B. In the other half of the project, to ensure capability of the model to question answer over images, Vision Encoder-Decoder architecture is used, it combines the capabilities of Vision Encoder i.e., ViT for processing images and Bidirectional Encoder Representations from Transformers i.e., BERT model to understand language while question answering.

The results are combines to make a wholesome project which can answer a questions on text as well as image data to do tasks like, context-based answering, completing a sentence, filling up missing details, solving queries and many more. The project can be used in various real-world domains like creating a fully capable chatbot for industry-specific use cases or creating an instructor bot etc.

1. Conclusion

Image Question answering Model:

This model successfully answer and complete the given task in context of the images give as input to the model. The generated output were semantically meaningful, coherent, and accurate in the desired context. Evalualtion metrices like ROUGE and BLEU give high level of accuracy between generated output and human thoughts.

Document-based Question Answering Model:

On asking question related to the document and URLs given as input to the model, the model generated contextually correct answers. The outputs were highly conciseness, reflecting the better understanding and extraction of information from different sources.

By leveraging the power of different Large Language Models , both the models demonstrate high level of capabilities in respective tasks. The model is highly scalable has capabilities of better performance in different use cases.

Further research and development in this area are essential to enhance model performance, scalability, and usability, paving the way for advanced multimodal applications with broader societal impact.

References

[1] A. Salaberria, G. Azkune, O. Lopez de Lacalle, A. Soroa, and E. Agirre, “Image captioning for effective use of language models in knowledge-based visual question answering,” *Expert Syst Appl*, vol. 212, Feb. 2023, doi: 10.1016/j.eswa.2022.118669.

[2] F. Ren and Y. Zhou, “CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering,” *IEEE Access*, vol. 8, pp. 50626–50636, 2020, doi: 10.1109/ACCESS.2020.2980024.

[3] Y. Bazi, M. M. Al Rahhal, L. Bashmal, and M. Zuair, “Vision–Language Model for Visual Question Answering in Medical Imagery,” *Bioengineering*, vol. 10, no. 3, Mar. 2023, doi: 10.3390/bioengineering10030380.

[4] Y. Li *et al.*, “Visual Question Generation as Dual Task of Visual Question Answering.”

[5] K. Kafle and C. Kanan, “An Analysis of Visual Question Answering Algorithms.”

[6] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, “BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.09936

[7] P. Ahir and H. Diwanji, “Knowledge Detection by Relevant Question and Image Attributes in Visual Question Answering.”

[8] A. Mostafa, H. Abbas, and M. I. Khalil, “Comparative Study of Visual Question Answering Algorithms,” in *Proceedings of ICCES 2020 - 2020 15th International Conference on Computer Engineering and Systems*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICCES51560.2020.9334686.

[9] A. Salaberria, G. Azkune, O. Lopez de Lacalle, A. Soroa, and E. Agirre, “Image captioning for effective use of language models in knowledge-based visual question answering,” *Expert Syst Appl*, vol. 212, Feb. 2023, doi: 10.1016/j.eswa.2022.118669.

[10] A. K. Gupta, “Survey of Visual Question Answering: Datasets and Techniques,” May 2017, [Online]. Available: http://arxiv.org/abs/1705.03865

[11] A. Agrawal *et al.*, “VQA: Visual Question Answering,” May 2015, [Online]. Available: http://arxiv.org/abs/1505.00468

[12] S. Lu *et al.*, “Improved Blending Attention Mechanism in Visual Question Answering,” *Computer Systems Science and Engineering*, vol. 47, no. 1, pp. 1149–1161, 2023, doi: 10.32604/csse.2023.038598.

[13] Z. He, Y. Li, and D. Zhang, “Transformer-based visual question answering model comparison,” *J Phys Conf Ser*, vol. 2646, no. 1, p. 012031, Dec. 2023, doi: 10.1088/1742-6596/2646/1/012031.

[14] D. Tang, N. Duan, T. Qin, Z. Yan, and M. Zhou, “Question Answering and Question Generation as Dual Tasks,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.02027

[15] L. Song, Z. Wang, and W. Hamza, “A Unified Query-based Generative Model for Question Generation and Question Answering,” Sep. 2017, [Online]. Available: http://arxiv.org/abs/1709.01058