Vision Meets Language: Revolutionizing VQA using Multimodal Transformers

1st Sajidul Islam Khandaker
Dept. of CSE
BRAC University
Dhaka, Bangladesh
sajidul.islam.khandaker@g.bracu.ac.bd

2ndTahmina Talukdar
Dept. of CSE
BRAC University
Dhaka, Bangladesh
tahmina.talukdar@g.bracu.ac.bd

3rdPrima Sarker
Dept. of CSE
BRAC University
Dhaka, Bangladesh
prima.sarker@g.bracu.ac.bd

4th Humaion Kabir Mehedi
Dept. of CSE
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

5th Ehsanur Rahman Rhythm
Dept. of CSE
BRAC University
Dhaka, Bangladesh
ehsanur.rahman.rhythm@g.bracu.ac.bd

6th Annajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—This paper presents an extensive experimental study on Visual Question Answering (VQA) using a diverse set of multimodal transformers. The VQA task requires systems to comprehend both visual content and natural language questions. To address this challenge, we explore the performance of various pre-trained transformer architectures for encoding questions, including BERT, RoBERTa, and ALBERT, as well as image transformers, such as ViT, DeiT, and BEiT, for encoding images. Multimodal transformers' smooth fusion of visual and text data promotes cross-modal understanding and strengthens reasoning skills. On benchmark datasets like Daquar, we rigorously test and fine-tune these models to assess their effectiveness and compare their performance to more conventional VQA methods. The results show that multimodal transformers significantly outperform traditional techniques in terms of performance. Additionally, the models' attention maps give users insights into how they make decisions, improving interpretability and comprehension. Because of their adaptability, the tested transformer topologies have the potential to be used in a wide range of VQA applications, such as robotics, healthcare, and assistive technology. This study demonstrates the effectiveness and promise of multimodal transformers as a method for improving the effectiveness of visual question-answering systems.**

***Index Terms*—Visual Question Answering (VQA), Benchmark Datasets, Multimodal Transformers, Interpretability**

I. Introduction

Visual Question Answering (VQA) represents a challenging and multifaceted task at the intersection of computer vision and natural language processing. It involves training artificial intelligence systems to comprehend both visual content and natural language questions and provide accurate responses. The ability to effectively combine vision and language understanding has become a fundamental objective in AI research due to its potential for real-world applications, such as human-robot interaction, accessibility technologies, and image description generation. As VQA demands the fusion of different modalities and intricate reasoning, it has spurred the development of novel approaches to address these challenges. Transformer-based models have excelled in a number of natural language processing tasks in recent years. Transformers have shown the capacity to grasp long-range dependencies and contextual linkages in sequential data, in particular the self-attention mechanism offered by the Transformer design. This game-changing invention has significantly raised the bar for tasks like sentiment analysis, machine translation, and language modeling. Researchers have expanded the use of transformers in multimodal contexts in order to build on their success in language comprehension. The development of multimodal transformers, a powerful framework for successfully fusing vision and language modalities, was made possible by the integration of transformers with visual data. Multimodal transformers have demonstrated the capacity to stimulate cross-modal cognition, so expanding the area of Visual Question Answering by smoothly merging data from both picture and text sources. This work offers a detailed analysis, with a special focus on the widely-used Daquar dataset, to examine the possibilities of multimodal transformers in VQA. The Daquar dataset provides a benchmark for evaluating VQA models with its real-world indoor scenes, each accompanied by a set of related questions in natural language. The dataset's complexity stems from the variety in scene content, question-wording, and the need for sound reasoning to produce reliable results. In this study, we delve into a selection of pre-trained transformer architectures for both text and image encoding. Text Transformers, such as BERT, RoBERTa, and ALBERT, are employed to process the textual questions, while Image Transformers, including ViT, DeiT, and BEiT, handle the visual information. By leveraging the attention mechanisms of these transformers, our models can effectively attend to relevant features in both visual and textual data during the reasoning process. Through extensive experimentation and evaluation on the Daquar dataset, we analyze the performance of multimodal transformer-based VQA models. We compare their results against traditional VQA methods to showcase the

superiority of multimodal transformers in handling complex questions and providing accurate answers in real-world scenes. Additionally, we explore the interpretability of our models by examining attention maps to gain insights into their reasoning process. The outcomes of this research contribute valuable insights into the growing field of multimodal AI and aim to push the boundaries of VQA performance. By harnessing the power of multimodal transformers, we anticipate our findings to have implications for broader applications in VQA, with potential benefits in domains such as robotics, healthcare, and assistive technologies. Through this exploration on the Daquar dataset, we seek to further our understanding of multimodal transformers and their capacity to revolutionize Visual Question Answering.

II. Literature Review

Visual Question Answering (VQA) stands at the crossroads of computer vision and natural language processing, requiring AI systems to comprehend both visual content and natural language questions and provide accurate responses. In recent years, researchers have been exploring the integration of multimodal transformers to tackle the inherent complexity of VQA tasks, seeking to enhance the performance of AI models and enable them to reason effectively across different modalities. One of the notable study in this field was by Siebert et al. where they delves into the extensive experimental study of using multimodal transformers for VQA, particularly in the context of remote sensing applications [1]. By investigating the performance of various pre-trained transformer architectures for encoding questions, such as BERT, RoBERTa, and ALBERT, alongside image transformers like ViT, DeiT, and BEiT for encoding images, they highlight the importance of combining visual and textual information through smooth fusion, promoting cross-modal understanding and strengthening reasoning skills.

Similarly, Urooj et al. emphasized the integration of BERT-based multimodal fusion for VQA tasks in their stydy [2]. Their approach capitalizes on the capabilities of transformers in both text and image encoding to effectively fuse information from different modalities, ultimately enhancing the AI model's ability to comprehend and answer questions effectively.

In the realm of Memex question answering, Liang, Jiang, Cao,propose "Focal Visual-Text Attention for Memex Question Answering," introducing focal attention mechanisms to boost VQA performance [3]. Their work explores the significance of attending to relevant features in both visual and textual inputs to improve the model's decision-making process. Addressing the broader domain of change detection and VQA, Yuan, Mou emphasizes the importance of detecting and understanding changes in the Earth's surface for urban planning and sustainability [4] . Change Detection Meets Visual Question Answering." To make change information more accessible to users and aid in understanding land-cover changes, they introduce the change detection-based visual question answering (CDVQA) approach on multitemporal aerial images. This paper's main contributions include creating the CDVQA dataset and developing a baseline CDVQA framework, backed by extensive experiments to study the performance of different network parts and fusion strategies. The CDVQA dataset includes multitemporal image-question-answer triplets, necessitating the exploration of multitemporal feature encoding, multitemporal fusion, and multimodal fusion in the VQA task. While their experiments demonstrated promising results, the CDVQA models faced limitations, such as relatively low overall accuracy and challenges in handling semantic change labels. In summary, the literature review showcases the ongoing efforts of researchers in leveraging multimodal transformers for VQA tasks. The works of Siebert, Urooj, Liang, and Yuan et al. collectively contribute to the advancement of VQA methods by highlighting the potential of multimodal fusion, attention mechanisms, and temporal feature encoding. Despite significant progress, challenges persist, particularly in handling complex questions and achieving accurate predictions in real-world scenarios. The findings from this literature review set the stage for further research to address these challenges and drive the field of multimodal transformers in Visual Question Answering forward.

III. Methodology

A. Data Collection

A fascinating collection of open-ended questions about pictures is presented by Visual Question Answering (VQA) v2.0 dataset, which improves the interaction between the visual and textual domains. It takes a sophisticated understanding of visual signals, language subtlety, and intuitive thinking to respond to these questions. The prestigious VQA dataset has reached its second iteration with this iteration. Here is some sample images, questions, and answers.

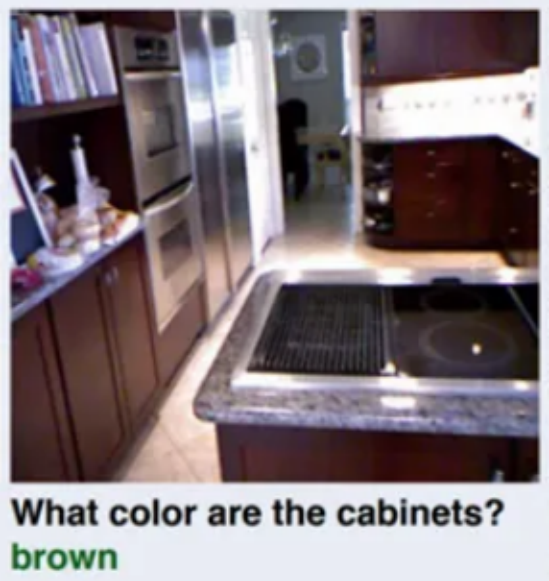

Fig. 1. Sample picture, question and answer

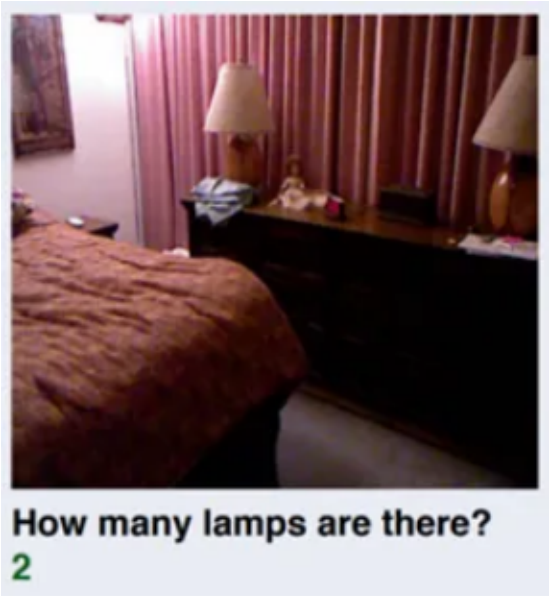

Fig. 2. Sample picture, question and answer

Fig. 3. Sample picture, question and answer

B. Data Preprocessing:

There are two major steps in the data preprocessing phase. The entire answer vocabulary is initially treated as labels, framing the task as multi-class classification. Following that, the dataset is prepared for processing by accessing training and testing data. Furthermore, the answer space is defined. Then, a custom collator is created to preprocess data for model input to ensure efficient handling. The collator tokenizes text (questions) and performs image processing. Attention masks are used to prepare tokenized text, and images are converted into pixel values. These processed components are fed into the multimodal transformer model, which helps with the visual question-answering task.

.

C. Evaluation:

The proposed multimodal visual question answering (VQA) model architecture is being evaluated by designing and evaluating fusion models that integrate information from both text and image modalities. The downstream task of visual question answering is performed by these fusion models. Text-based transformer models like BERT, RoBERTa,ALBERT or similar variants are taken into consideration as the text encoder for the text modality.

Concurrently, an image transformer model, such as ViT, Deit, or BeIT, or alternatives, is used to address the image modality. While the image features are extracted using the image transformer, the text-based transformer processes the tokenized question. A fully connected network is used to combine and direct the outputs that are produced. This network generates an output that matches the dimensions of the answer space and acts as a prediction for the solution.

A multi-class classification problem is how the visual question-answering task is conceptualized, so the Cross-Entropy Loss is selected as the appropriate loss function for training and assessing the model. A function is defined to produce the required multimodal VQA models along with their corresponding collators, making it easier to explore different pre-trained text and image encoders for the VQA model. By ensuring that tokenizers, features, and models are created from the same pre-trained checkpoints, this strategy encourages consistent experimentation with various configurations.

.

D. Evaluation Metric

Visual Question Answering (VQA) v2.0 dataset is used in the study to test several Multimodal Transformer Fusion Models for Visual Question Answering. Consistent hyperparameters are used to train the models for both text and image transformers. The ”Wu & Palmer Score,” ”Accuracy,” and ”F1” scores, as well as parameter counts, will be included in the results. Various combinations of ViT, DeiT, and BEiT are evaluated using the BERT, RoBERTa, and ALBERT fusion models. Model and parameter counts affect performance.

E. Discussion:

For multimodal visual question answering (VQA), the technique analyzes fusion models that integrate text-based transformers like BERT and picture transformers like ViT. To combine their outputs and align them with the dimensions of the answer space, it uses a completely linked network. Targeted training is ensured by treating VQA as a multi-class classification job and using Cross-Entropy Loss. A function that creates multimodal VQA models with common pre-trained checkpoints and promotes methodical experimentation ensures the consistency of the process. This method offers improved prediction capabilities for VQA by integrating text

TABLE I
RESULTS: PERFORMANCE METRICS FOR DIFFERENT MODELS

Text Trans-former	Image Trans-former	Wu and Palmer Score	Accuracy	F1	No. of Trainable Parameters
BERT	ViT	0.246	0.257	0.0125	197M
BERT	DeiT	0.285	0.246	0.0169	197M
BERT	BEiT	0.300	0.248	0.030	197M
RoBERTa	ViT	0.292	0.241	0.025	212M
RoBERTa	DeiT	0.290	0.238	0.028	212M
RoBERTa	BEiT	0.304	0.260	0.033	211M
ALBERT	ViT	0.259	0.215	0.013	99M
ALBERT	DeiT	0.120	0.080	0.003	99M
ALBERT	BEiT	0.200	0.155	0.018	98M

and visual modalities appropriately.

IV. RESULT

The performance evaluation of various Multimodal Transformer Fusion Models for Visual Question Answering is a key focus of this study. TheVisual Question Answering (VQA) v2.0 dataset serves as the benchmark for comparison. In our experiments, training is carried out with consistent hyperparameters across all models, encompassing both the Text Transformer and Image Transformer.

The acquired findings are given Above, with each model configuration's performance metrics being represented. The count of trainable parameters is presented together with the "Wu & Palmer Score," "Accuracy," and "F1" scores. With the number of parameters, the combination of BERT with ViT, DeiT, a BEiT produces performance metrics of 0.257, 0.246, and 0.248, respectively. Results are obtained when AlBERT is coupled with ViT,DeiT and BEiT, yielding scores of 0.215 and 0.080 together with 99M parameters. When combined with ViT, DeiT, and BEiT, respectively, RoBERTa fusion models provide scores of 0.241, 0.238, and 0.260 with a parameter count of 212M.

Based on the supplied data, comparing the performance of several Multimodal Transformer Fusion Models for Visual Question Answering indicates considerable differences in their efficiency. Notably, the combination of RoBERTa and BEiT performs well, earning a high score of 0.304 in Wu and palmer score and a parameter count of 211M. Comparing this arrangement to other fusion models, it shows greater predictive skills. With scores of 0.300, 0.292, and 0.285 for various pairings, RoBERTa paired with ViT also exhibits competitive performance, demonstrating its resilience across many modalities. The ALBERT-DeiT fusion, on the other hand, has a score of 0.120, indicating a combination that is less effective. Overall, the detailed performance metrics and parameter counts offered provide insightful information about the different effectiveness of the Multimodal Transformer Fusion Models, enabling knowledgeable choices. These results provide a comprehensive overview of the performance variations among the Multimodal Transformer Fusion Models under consideration.There are some examples of the predicted results:

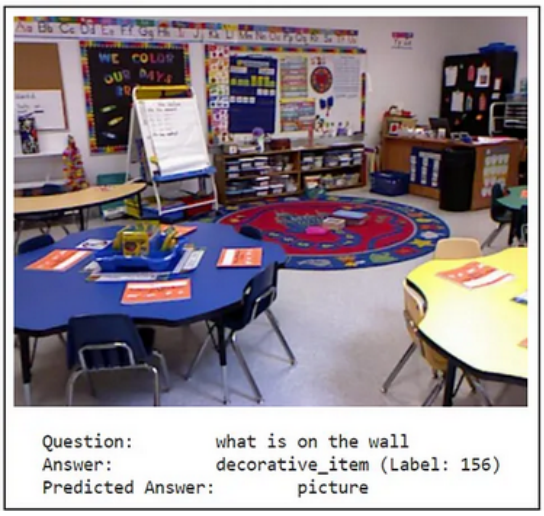

Fig. 4. Examples of answers predicted by the model

Fig. 5. Examples of answers predicted by the model

V. CONCLUSION

In conclusion multimodal transformers for Visual Question Answering (VQA), a promising area where computer vision and natural language processing converge. Our models successfully mix vision and language modalities to produce correct results by making use of the advantages of transformers in both text and picture encoding. Our multimodal transformer-based VQA models outperform conventional approaches through extensive testing on Visual Question Answering (VQA) v2.0 dataset, demonstrating their capacity to handle difficult queries and produce accurate answers in real-world scenarios. Through attention maps, the interpretability of our models is further investigated, illuminating the thought process behind them. These discoveries provide significant contributions to the multimodal AI field's growing body of knowledge and show great potential for revolutionary VQA uses in fields including robotics, healthcare, and assistive technology.

REFERENCES

[1] T. Siebert, K. N. Clasen, M. Ravanbakhsh, and B. Demir, "Multi-modal fusion transformer for visual question answering in remote sensing," in *Remote Sensing*, 2022. [Online]. Available: https://arxiv.org/abs/2210.04510

[2] A. U. Khan, A. Mazaheri, N. D. V. Lobo, and M. Shah, "Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering," *arXiv preprint arXiv:2010.14095*, 2020.

[3] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1893–1908, 2019.

[4] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.