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
ID:20301190

2ndTahmina Talukdar
Dept. of CSE
BRAC University
Dhaka, Bangladesh
tahmina.talukdar@g.bracu.ac.bd
ID:20301412

3rd Prima Sarker
Dept. of CSE
BRAC University
Dhaka, Bangladesh
prima.sarker@g.bracu.ac.bd
ID:20301204

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

5thAnnajiat 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 papers contributing to this area of research is "Multi-Modal Fusion Transformer for Visual Question Answering in Remote Sensing" by Siebert, Clasen, Ravanbakhsh, and Demir. Their work delves into the extensive experimental study of using multimodal transformers for VQA, particularly in the context of remote sensing applications. 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, Mazaheri, Da vitoria lobo, and Shah present "MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering," emphasizing the integration of BERT-based multimodal fusion for VQA tasks. 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, Kalantidis, Li, and Hauptmann propose "Focal Visual-Text Attention for Memex Question Answering," introducing focal attention mechanisms to boost VQA performance. 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, Xiong, and Zhu present "Change Detection Meets Visual Question Answering." Their work emphasizes the importance of detecting and understanding changes in the Earth's surface for urban planning and sustainability. 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. 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 the Daquar 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