iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60786

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Visual Mind: Visual Question Answering (VQA) with CLIP Model

Ruchita Sonawale[1], Arshia Shaikh[2]
Computer Engineering, University of Mumbai,

Abstract: ***This paper proposes a Visual Question Answering (VQA) problem using CLIP models. The proposed approach suggests an enhanced VQA-CLIP model with additional layers for better computational performance. VQA is an increasingly important task that aims to answer open-ended questions based on images. This task has numerous applications in various fields such as medicine, education, and surveillance. The VizWiz dataset, specifically designed to assist visually impaired individuals, consists of image/question pairs along with 10 answers per question, recorded by blind participants in a natural setting. The task involves predicting answers to questions and determining when a question is unanswerable. In this study, we will utilize the VizWiz dataset and employ the CLIP model with an additional linear layer, a multimodal, zero-shot model known for its efficiency in processing image and text data. By leveraging the unique capabilities of CLIP, and benchmarked against state-of-the-art approaches. Results indicate a competitive or better performance of the VQA model.***
Index Terms: Visual Question Answering (VQA), CLIP model, VizWiz dataset.

I. INTRODUCTION

The Visual Question Answering (VQA) paradigm represents a fascinating convergence of computer vision and natural language processing, aimed at imbuing machines with the ability to comprehend and respond to questions about images in natural language [1], [2]. As visual content becomes increasingly prevalent in our digital landscape, ranging from social media platforms to applications in autonomous vehicles, the demand for intelligent systems capable of seamlessly integrating visual information with textual input continues to grow. This interdisciplinary field seeks to bridge the gap between visual and textual modalities, laying the foundation for machines to engage in nuanced reasoning about images through natural language understanding.

In recent years, the exploration of VQA has been characterized by a quest to overcome the inherent challenges posed by the fusion of visual and textual information [1], [3]. Early approaches primarily focused on simplistic feature fusion techniques, such as simple concatenation or element-wise products, to combine image and text-based features. However, as the complexity of VQA tasks increases, there is a pressing need to explore more sophisticated fusion methods to unlock the full potential of these multimodal interactions.

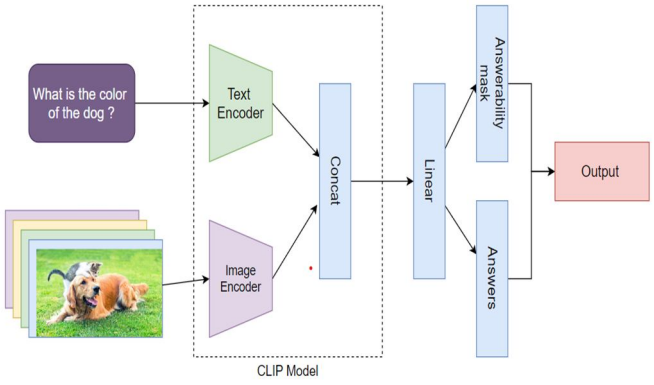


Fig. 1 An overview of the proposed framework of VQA with CLIP

Our research endeavors to address this challenge by delving into advanced fusion techniques, aiming to elevate the performance and capabilities of VQA systems. By leveraging empirical evaluations on benchmark datasets, we seek to demonstrate the efficacy of our proposed methods in tackling the intricacies of VQA tasks.

©IJRASET: All Rights are Reserved

Central to our approach is the exploration of attention mechanisms across different modalities, recognizing their pivotal role in capturing and leveraging contextual information for accurate question answering.

Indeed, the success of VQA models hinges on their ability to effectively exploit attention mechanisms, both within individual modalities and across modalities [6], [8]. By incorporating question-guided attention mechanisms on images and exploring reciprocal attention from images to questions, our research aims to enrich the depth of understanding and reasoning capabilities of VQA systems.

II. LITERATURE SURVEY

Visual Question Answering (VQA) stands at the crossroads of computer vision and natural language processing, challenging models to understand and respond to queries based on both visual and textual inputs. The nascent stages of VQA research saw the emergence of handcrafted feature engineering and rule-based methodologies. These early endeavors laid the groundwork for subsequent advancements by introducing fundamental datasets and baseline models. Particularly noteworthy is the work of Antol et al. (2015) [14], who pioneered the VQA dataset and proposed an initial baseline model combining bag-of-words features with convolutional neural network (CNN) representations.

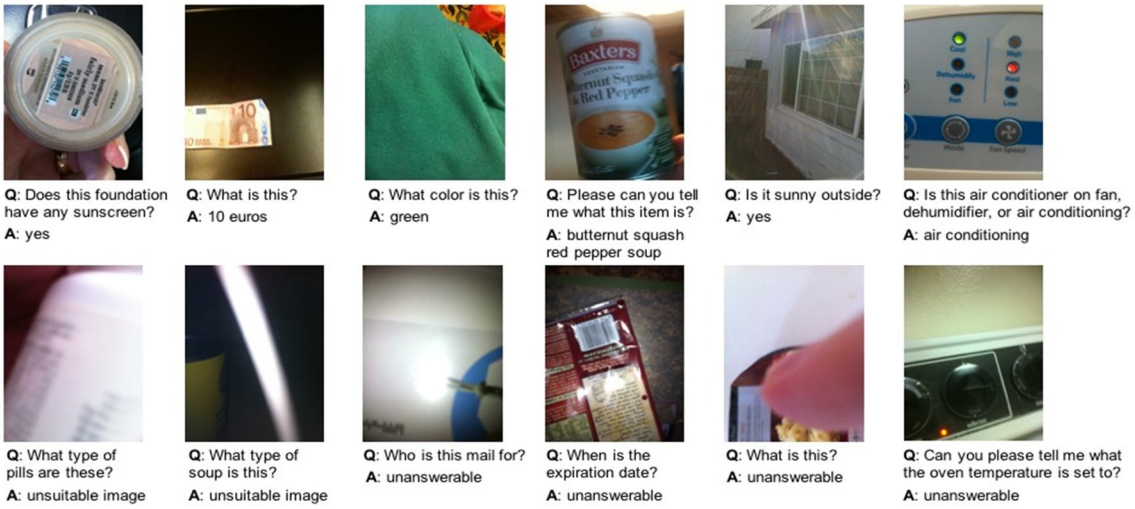


Fig 2. Vizwiz dataset sample images and question pairs

The advent of deep learning heralded a paradigm shift in VQA, as researchers began exploring end-to-end trainable models capable of learning from raw data. This transition birthed innovative techniques such as the Multimodal Compact Bilinear pooling (MCB) method proposed by Malinowski et al. (2015) [15], showcasing the efficacy of deep learning in tackling VQA tasks.

Multimodal fusion techniques emerged as a focal point in VQA research, aiming to effectively integrate visual and textual information for improved performance. Models like the Hierarchical Co-Attention (HieCoAtt) (Lu et al., 2016) [16] and Multimodal Compact Bilinear Attention Networks (MCBAN) (Fukui et al., 2016) [17] introduced simultaneous attention mechanisms, enhancing the model's ability to focus on relevant features.

Attention mechanisms played a pivotal role in advancing VQA models by enabling selective focus on salient visual and textual cues. Notable contributions include the Stacked Attention Networks (SAN) (Yang et al., 2016) [18] and the Bottom-Up and Top-Down Attention model (BUTD) (Anderson et al., 2018) [19], which demonstrated significant improvements in performance through sophisticated attention mechanisms.

Recent developments in VQA research have witnessed the rise of transformer-based architectures, leveraging the power of transformers to capture long-range dependencies effectively. Models such as VisualBERT (Tan et al., 2019) [20], Vision Transformer (ViT) (Dosovitskiy et al., 2021) [21], and Data-efficient Image Transformer (DeiT) (Touvron et al., 2020) [22] have pushed the boundaries of VQA performance, achieving state-of-the-art results on various benchmarks.

Despite the progress, several challenges persist in VQA research, including multimodal ambiguities, dataset biases, and model interpretability. Addressing these challenges and exploring novel architectures, self-supervised learning techniques, and robust evaluation metrics are imperative for the continued advancement.

In conclusion, this comprehensive literature survey provides a detailed exploration of the evolution of VQA models, highlighting key contributions, advancements, and challenges. By synthesizing insights from existing research, we aim to inspire further innovation and progress in the dynamic field of visual question answering.

III. EXPERIMENTS

A. Dataset - VIzWiz

As a team, the Visual Question Answering (VQA) task we are tackling involves developing AI models that can answer open-ended questions about images. While there are several popular VQA datasets like VQA v2 and OK-VQA, the specific dataset we will be utilizing is called VizWiz.

VizWiz is a unique VQA dataset that was created to help address the technological needs of visually impaired people. It contains images and questions that were sourced directly from blind users who took pictures and recorded spoken questions about them.

Some key points about the VizWiz dataset that we want to highlight:

1) It contains 20,500 image-question pairs in total.

2) For each image, there is one corresponding question asked by a blind person, as well as 10 different answers provided by crowdsourced workers.

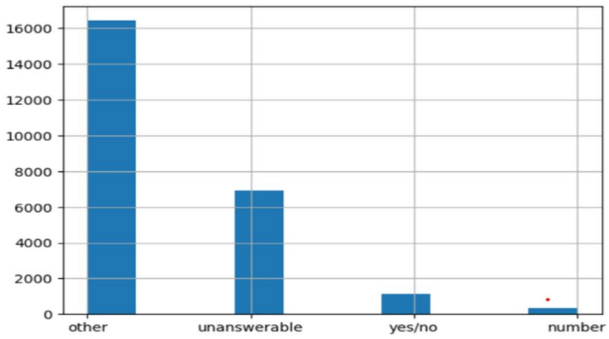


Fig.3 The histogram shows the answer type in the dataset

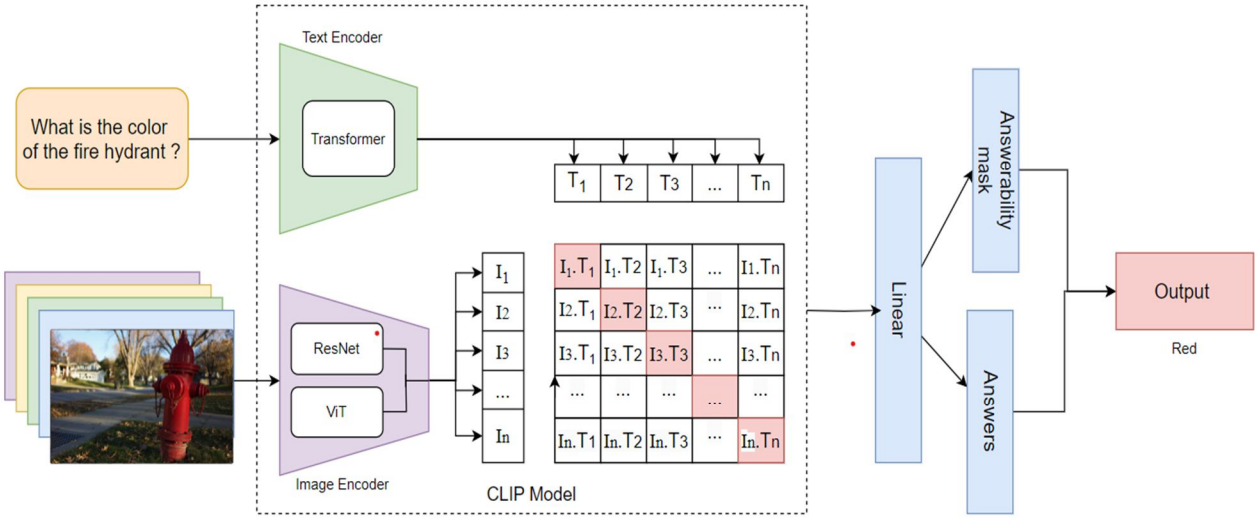


Fig. 3 Architecture of the proposed framework of VQA with CLIP

3) The dataset is split into:
- 20,523 image-question pairs for training
- 205,230 answer/confidence score pairs for training
- 4,319 image-question pairs for validation
- 43,190 answer/confidence score pairs for validation

4) The questions were asked by blind people in a natural setting about images they captured, making this a realistic and challenging dataset for our team to work with.

5) In addition to predicting the answer, one of the tasks we need to tackle is predicting whether a given question is unanswerable based on the image.

In summary, VizWiz provides a unique real-world VQA dataset sourced from blind users, with the goal of developing assistive technologies to help this population. The multiple reference answers and answerability prediction make it a particularly challenging dataset for us as a team.

IV. PROPOSED APPROACH

We have chosen to utilize the CLIP (Contrastive Language-Image Pre-training) model for our Visual Question Answering (VQA) task on the VizWiz dataset. The primary reasons behind this choice are the unique advantages and capabilities offered by CLIP.

CLIP is an open-source, multimodal, and zero-shot model, which means it can effectively process and understand both image and text data without being explicitly trained on a specific task. Given an image and text descriptions, CLIP can predict the most relevant text description for that image, leveraging its pre-training on a vast amount of image-text pairs.

One of the key advantages of using CLIP for our VQA task is its ability to handle multimodal inputs, which is essential for answering questions based on both visual and textual information. Additionally, CLIP's zero-shot capabilities allow us to leverage its pre-trained knowledge, enabling efficient training and adaptation to the VizWiz dataset without requiring extensive computational resources.

CLIP offers several model variants, each with varying architectures and capacities, such as RN50, RN101, RN50x4, RN50x16, RN50x64 and ViT-B/32. These variants differ in their backbone architectures (ResNet or Vision Transformer), model sizes, and input image resolutions, allowing us to choose the most suitable configuration based on our computational constraints and performance requirements.

Our approach involves utilizing both the image and text encoders of the CLIP model. The resulting features from these encoders are concatenated and passed through additional linear layers with layer normalization and a high dropout value (0.5) to capture the multimodal representations effectively. The answer types and final answers are predicted using separate linear layers, as illustrated in Figure 2 of our methodology.

The image size for the visual encoder is set to 448x448 for the RN50x64 variant and 336x336 for the ViT-L/14@336px variant. During training, we employ cross-entropy loss and rotation as an image augmentation technique to enhance the model's robustness and generalization capabilities.

Notably, we train only the additional linear classifier components while keeping the pre-trained CLIP model weights frozen. This approach allows us to leverage CLIP's pre-trained knowledge and representations, enabling fast and efficient training without requiring extensive computational resources or large-scale fine-tuning on the VizWiz dataset.

By adopting the CLIP model and our proposed methodology, we aim to leverage its multimodal capabilities, zero-shot learning, and efficient training process to develop an effective solution for the Visual Question Answering task on the VizWiz dataset. This approach holds promise for addressing the technological needs of visually impaired individuals while advancing the field of assistive technologies.

V. RESULTS

The table presents the performance results of three different VQA models: ResNet-50 (RN50), ResNet-101 (RN101), and Vision Transformer with a base architecture (ViT-B/32).

Model	VQA (Acc.)	Answerability (Acc.)

	train	val	test	train	val	test
RN50	70.48%	66.80%	60.56%	97.8913 %	94.2264 %	95.3270 %
RN101	78.8711 %	68.9674 %	62.0448 %	97.9239 %	94.4273 %	95.9179 %
ViT-B/32	90.92%	72.10%	65.76%	97.43%	93.98%	95.56%

Across all models, the Vision Transformer (ViT-B/32) consistently outperforms the ResNet models in terms of VQA accuracy across all datasets, including the training, validation, and test sets. This indicates the potential superiority of transformer-based architectures in handling visual question answering tasks, possibly due to their ability to capture long-range dependencies in the data more effectively than convolutional neural networks.

However, while the ViT-B/32 model demonstrates the highest VQA accuracy, there is still a noticeable performance gap between the training and test sets for all models. This gap suggests potential issues with overfitting, where the models may have memorized patterns specific to the training data but struggle to generalize to unseen data. Addressing this gap is crucial for ensuring the models' robustness and reliability in real-world applications.

Additionally, the table includes metrics for answerability accuracy, which remain consistently high across all models and datasets. This indicates that the models perform well in determining whether an answer can be provided for a given question-image pair, regardless of the specific architecture used. High answerability accuracy is essential for effectively filtering out unanswerable questions and improving the overall user experience of VQA systems.

The results also highlight the trade-off between model complexity and performance. As the model complexity increases, from ResNet-50 to ResNet-101 to ViT-B/32, there is a corresponding improvement in VQA accuracy. However, this improvement comes with increased computational costs, both in terms of training time and inference complexity. Therefore, it's essential to carefully balance model complexity with practical considerations when designing VQA systems for real-world deployment.

VI. FUTURE SCOPE

Real-time Interaction: The incorporation of real-time interaction capabilities into VQA systems could significantly enhance their practical utility. Enabling users to ask questions and receive answers in real-time, especially in dynamic environments, would contribute to a more seamless and responsive user experience.

Semantic Understanding and Context Awareness: Future research could focus on improving the semantic understanding of visual content and context awareness in VQA systems. This involves recognizing subtle details, spatial relationships, and temporal elements in images, leading to more contextually relevant responses. Integration with Assistive Technologies: Exploring synergies with existing and emerging assistive technologies offers potential for creating comprehensive solutions. Integrating VQA capabilities with wearable devices, voice assistants, or tactile feedback systems could augment the overall user experience for visually impaired individuals.

VII. APPLICATION

Assistive technology for the visually impaired: VQA systems can help individuals with visual impairments find answers to questions about their environment, allowing them to be more independent in everyday functioning in various fields.

Educational tools: VQA can be incorporated into educational strategies to create interactive learning environments where students can ask questions about visual materials such as pictures or drawings, and receive immediate answers and explanations.

Customer service chatbots: Companies can use VQA-powered chatbots that can help customers with product-related questions, by analyzing images and answering questions about products, services, and problems.

Medical Image Analysis: In the medical field, the VQA can help radiologists and healthcare professionals interpret medical images by answering questions about anatomy, abnormalities, and diagnostic findings will be raised

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Smart Home Devices: VQA can enhance the functionality of smart home devices by allowing users to ask questions about their home environment, such as finding objects, recognizing faces, or monitoring activities.
E-Commerce Product Recommendations: E-commerce platforms can use VQA to improve product recommendations by understanding the questions users ask related to visual quality, preferences, and compatibility with existing purchases they will understand it.
Legal and Forensic Analysis: VQA systems can assist legal professionals and forensic investigators in analyzing visual evidence, such as crime scene photos or surveillance footage. Lawyers, judges, and investigators can ask questions about details or anomalies in images, helping to build stronger cases or uncover important insights during investigations

VIII. CONCLUSION

In conclusion, the use of the CLIP model in the context of Visual Question Answering (VQA) using the VizWiz dataset has shown promising results. By leveraging large pre-training data, and multi-modality of CLIP, we have demonstrated the model's capability as an excellent feature extractor for this task. The incorporation of the text Transformer from CLIP, alongside the image encoder, has led to improved performance in predicting answers to visual questions. Our proposed challenge not only aims to address the task of answering visual questions for blind individuals but also serves as an opportunity to develop assistive technologies that eliminate accessibility barriers for this population.

IX. APPENDIX

A. *VizWiz Dataset Description*

Description: The VizWiz dataset contains image/question pairs with 10 responses per question, recorded by visually impaired individuals.
Purpose: Designed to assist visually impaired individuals, the data set presents challenges to the VQA algorithm.

B. *Overview of the methodology*

Summary: A CLIP model with an additional linear layer was used for VQA projects on the VizWiz dataset.
Application: Image and text encoders, trained linear classifiers, and implementation of post vocabulary selection.

C. *Good building materials*

Description: CLIP-based VQA modeling with integrated image and text encoder, linear layer, and subsequent type prediction.
Features: Leverage CLIP OCR capabilities, pre-training data, and multiple capabilities for effective VQA.

D. *Summary of Outcomes*

Conclusions: The VizWiz data set provided promising results demonstrating the effectiveness of the CLIP-based model for VQA.
Performance implications: A model for accuracy, efficiency, and robustness in predicting responses to visual questions.

E. *The future*

Opportunities: Explored potential improvements including real-time communication, logic understanding and cross-domain optimization.
Applications: Potential applications in assistive technology, education, customer service, medical image analysis, and more were identified.

X. ACKNOWLEDGMENT

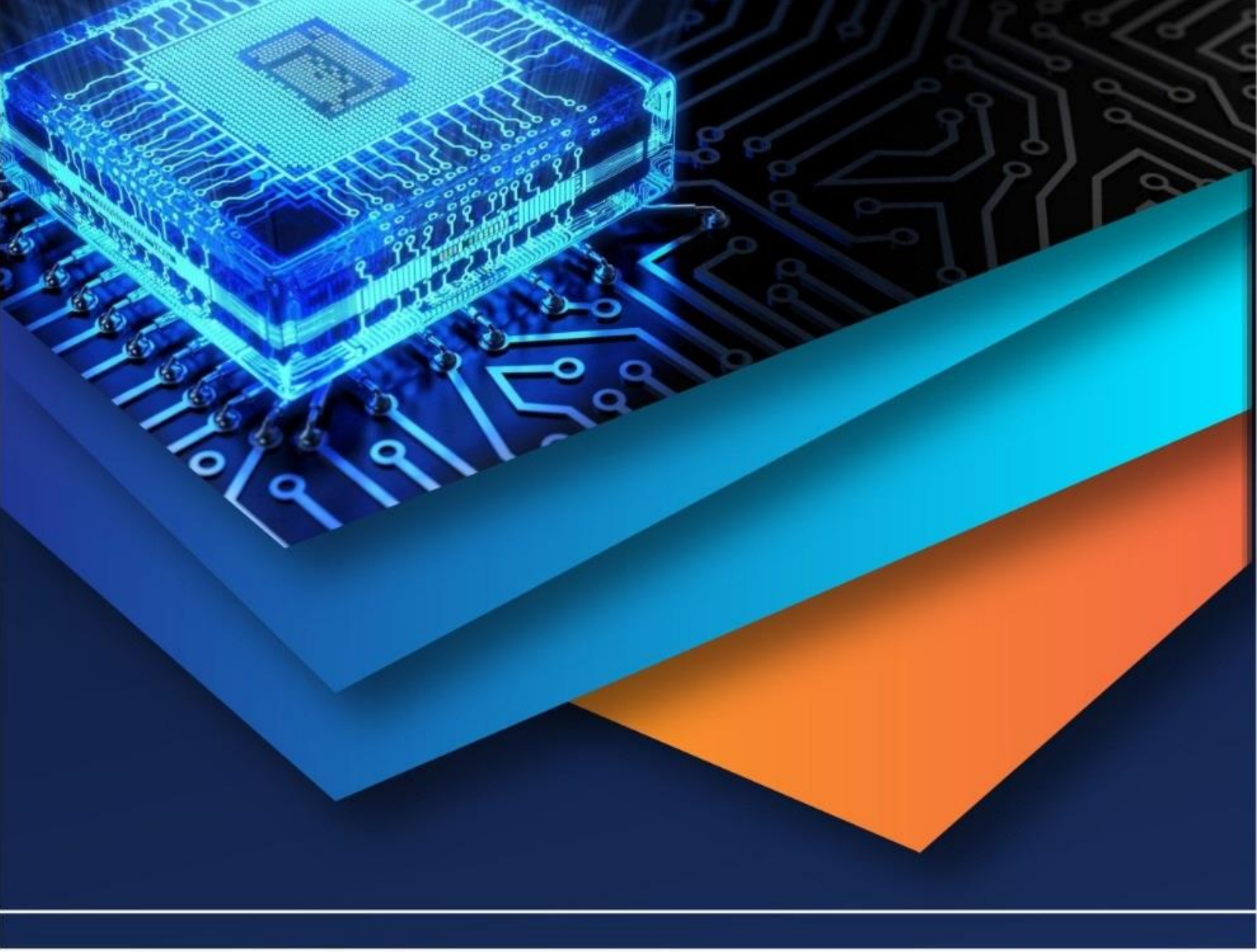
We would like to express our sincere thanks to all those who contributed to the completion of this paper. We express our sincere appreciation to Prof. Subodh, Project Guide, for their guidance, encouragement, and valuable expertise throughout this study. Their insightful feedback and unwavering support have been instrumental in shaping the direction of our research.
We are also very grateful to the Principal and Teachers of Department of Computer Engineering at Datta Meghe College of Engineering for providing us with necessary facilities and facilities to conduct our research successfully. Special thanks to our colleagues and fellow students who provided valuable insights and constructive criticism during the discussions and comments. Their feedback has improved the quality of our work and contributed to its overall success.

We would like to thank the researchers and developers of the CLIP model and the VizWiz data set for making their products explicit, which greatly facilitated our research efforts. Finally we express our sincere appreciation to our families and friends for their unwavering support, understanding and encouragement throughout this endeavor.
This research would not have been possible without all the collective efforts and contributions described above. Thank you for being a part of this journey.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proceedings of the IEE
[2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," International Journal of Computer Vision, vol. 123, no. 1, pp. 4–31, May
[3] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4613–4621.
[4] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21– 29.
[5] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," arXiv preprint arXiv:1704.03162, 2017.
[6] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances In Neural Information Processing Systems, 2016, pp. 289–297.
[7] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question guided spatial attention for visual question answering," in European Conference on Computer Vision. Springer, 2016, pp. 451–466.
[8] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in Proceedings of the IEEE Conference
[9] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
[10] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
[11] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
[12] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," IEEE Trans. Antennas Propagat., to be published.
[13] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),"
[14] IEEE J. Quantum Electron..
[15] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425-2433).
[16] Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1-9).
[17] Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In Advances in Neural Information Processing Systems (pp. 289-297).
[18] Fukui, A., Park, D. H., Yang, D., Rohrbach, M., Darrell, T., & Rohrbach, A. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In Advances in Neural Information Processing Systems (pp. 317-325).
[19] Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 21-29).
[20] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6077-6086).
[21] Tan, H., Xie, C., Li, S., Shen, X., & Zhou, X. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8132-8141).
[22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[23] Touvron, H., Caron, M., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2020). Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)