Ppt

Repo

Numbers data

Diagrams

Data Preparation

Literature , issue solving with this

Latex convert

Humanize

Plagiarism

Presentation should contain:

- Introduction to domain, topic

- literature survey ( atleast recent 15-20 papers) in tabular form.

- research gaps and challenges

- problem statement

- objectives ( 4 - 5 objectives for project)

- block diagram

- timeline chart of work done and to be done

Literature

# Literature Review

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.

#### 2.2.1. LXMERT

LXMERT (Learning Cross-Modality Encoder Representations from Transformers) is a prominent model in the VQA domain. It employs separate encoders for language and vision, followed by cross-modality encoders that enable interaction between the two modalities. This architecture allows LXMERT to effectively capture the complex relationships between visual elements and textual queries, resulting in improved VQA performance.

## 3. Key Studies

[1]

Tan and Bansal in 2019 introduced the LXMERT model, demonstrating significant improvements in VQA tasks by effectively integrating visual and textual information. It includes three types of encoders: object-relationship, language, and cross-modality. LXMERT is evaluated on popular datasets such as VQA, GQA, and NLVR2, outperforming previous state-of-the-art methods. This paper highlights LXMERT has superior performance on cross-modality tasks compared to previous BERT-like models, which focus solely on language.

[2]

VisualBERT model integrates BERT with image features extracted from object detectors, enabling it to jointly process text and image inputs using Transformers.It uses transformer architecture where image regions and textual inputs are treated as tokens and passed through self attention layers. The model is pre-trained on COCO dataset. VisualBERT goes through three training stages: task-agnostic pre-training on image-caption pairs, task-specific pre-training, and fine-tuning for specific tasks. VisualBERT provides a simple yet effective baseline for vision-and-language tasks, leveraging BERT’s language modeling capabilities

[3]

The acronym ViLBERT stands for Vision-and-Language BERT is a model that combines verbal and visual information to learn task-agnostic representations. By extending the BERT architecture, it creates a two-stream paradigm in which distinct streams interact through co-attentional transformer layers as they process textual and visual inputs. After pretraining the model on two proxy tasks from the Conceptual Captions dataset, it is optimized for vision and language tasks such as visual question answering, visual commonsense reasoning, and picture retrieval. In several tasks, ViLBERT outperforms current task-specific models by achieving state-of-the-art outcomes.

[4]

This paper explores the attentional gap between humans and machines to improve Visual Question Answering models for driving scenarios. The suggested method incorporates a filter to improve the VQA models' attention mechanism, giving driving-specific things like roads, cars, and signs priority. By utilizing the LXMERT model as an example, the filter enhances the precision of answers to queries about driving. Evaluations show that this approach enhances performance and aligns the model’s attention closer to human observation patterns.

[5]

This paper introduces Pixel-BERT, a unique deep multi-modal transformer architecture intended to improve text alignment with image pixels. Through the use of image-sentence pairs, Pixel-BERT improves the semantic links between language and visual by getting around the restrictions of region-based visual features. Experiments show that Pixel-BERT achieves state-of-the-art performance across a variety of vision-language tasks through thorough pre-training using datasets such as Visual Genome and MS-COCO, improving results like Visual Question Answering (VQA) by 2.17 points compared to previous models

[6]

**Large language models (LLMs) and multi-label learning of medical images are combined in the study Med-VQA to introduce the The Image to Label to Answer (ITLTA) framework, which addresses the problems of data shortages and complexity in Medical Visual Question Answering. The system minimizes deployment costs, improves interpretability and resilience, and eliminates the need for massive amounts of training data, thereby enabling zero-shot learning. Experiments conducted on the VQA-Med 2019 dataset demonstrate that the proposed method performs better than existing approaches, hence promoting the application of Med-VQA in clinical settings.**

**[7]**

**This paper gives us an innovative approach to Document Visual Question Answering (DocVQA) by extending it to multi-page documents through the introduction of the MP-DocVQA dataset. It also proposes Hi-VT5, a hierarchical multimodal transformer model capable of efficiently processing multi-page documents while maintaining explainability through page identification. Experimental results demonstrate that Hi-VT5 easily outperforms existing methods, validates the need for addressing multi-page contexts in DocVQA.**

### 3.4. PyTorch and Hugging Face in VQA

PyTorch has become a preferred framework for implementing deep learning models due to its dynamic computation graph and ease of use. Hugging Face’s Transformers library offers pre-trained models and tools that streamline the development of transformer-based architectures. Studies leveraging these tools, such as those by Wolf et al. (2020), have demonstrated accelerated model development and improved performance through efficient fine-tuning of pre-trained models.

## 4. Current Trends

### 4.1. 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 (Chen et al., 2020) and ViLT (Kim et al., 2021) have pushed the boundaries of VQA by enhancing cross-modal understanding and reasoning capabilities.

### 4.2. 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.

### 4.3. 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.

### 4.4. Explainability and Interpretability

As VQA models become more complex, there is a growing emphasis on making their decision-making processes transparent. Techniques like attention visualization and gradient-based methods are being utilized to interpret how models arrive at specific answers, fostering trust and facilitating debugging.

## 5. Gaps and Challenges

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.

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.

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.

### 5.4. 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.

## 6. Conclusion

The field of Visual Question Answering has made significant strides, particularly with the integration of transformer-based architectures like LXMERT, which enhance cross-modal understanding and reasoning. Datasets such as VQA v2.0 and Visual Genome have provided robust benchmarks for evaluating model performance. Current trends emphasize multimodal transformers, advanced attention mechanisms, data augmentation, and model interpretability, all contributing to the advancement of VQA technologies.

However, challenges related to dataset diversity, model generalizability, computational demands, and bias persist. Addressing these gaps is essential for the development of more robust, efficient, and fair VQA systems. Your project, which leverages transformers, PyTorch, and comprehensive datasets, is well-positioned to contribute to this evolving landscape by potentially exploring novel architectures, improving model efficiency, or enhancing dataset diversity.

---

## References

\*(Note: Below are example references. Please ensure to replace them with actual sources relevant to your project.)\*

- Chen, J., Li, L., Yu, K., & Sun, M. (2020). UNITER: Universal Image-Text Representation Learning. \*European Conference on Computer Vision (ECCV)\*.

- Goyal, Y., Khamzin, A., Parikh, D., & Batra, D. (2017). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. \*Proceedings of the IEEE International Conference on Computer Vision (ICCV)\*.

- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. \*International Journal of Computer Vision (IJCV)\*.

- Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. \*Advances in Neural Information Processing Systems (NeurIPS)\*.

- Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. \*arXiv preprint arXiv:1908.07490\*.

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. \*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations\*.

---

## Additional Tips for Writing Your Literature Review

1. \*\*Critical Analysis\*\*: Beyond summarizing existing work, critically analyze the strengths and weaknesses of different studies. Discuss how they contribute to the field and where they fall short.

2. \*\*Thematic Organization\*\*: Organize your review thematically rather than chronologically to provide a clearer narrative of the research landscape.

3. \*\*Synthesis Over Summary\*\*: Synthesize information from various sources to highlight patterns, relationships, and overarching themes rather than merely summarizing each paper individually.

4. \*\*Use of Visuals\*\*: Incorporate tables, charts, or diagrams to compare different models, datasets, or methodologies. Visual aids can enhance the clarity and impact of your review.

5. \*\*Stay Current\*\*: Ensure that your literature review includes the most recent studies and developments up to your knowledge cutoff in October 2023 to maintain relevance.

6. \*\*Proper Referencing\*\*: Adhere to a consistent citation style (e.g., APA, MLA) and ensure all sources are accurately referenced to avoid plagiarism.

7. \*\*Iterative Refinement\*\*: Writing a literature review is an iterative process. Continuously update and refine your review as you discover new sources or as your understanding of the topic deepens.

---

Feel free to modify and expand upon this example to better fit the specific requirements and scope of your project. If you need further assistance with any section or additional examples, let me know!

**Abstract**

ViLBERT introduces a novel two-stream model that learns joint, task-agnostic representations of visual and linguistic data using a multi-modal co-attentional mechanism. By pretraining on a large dataset, it drastically improves performance on various vision-and-language tasks like visual question answering and caption-based retrieval, surpassing existing models that rely on task-specific training. This approach shifts the paradigm by treating grounding between vision and language as pretrainable and transferable, leading to state-of-the-art results across multiple benchmarks.

**Key Points**

* ViLBERT extends the BERT architecture into a two-stream model to process visual and textual data separately while enabling interaction through co-attentional layers.
* The model is pretrained on the Conceptual Captions dataset using proxy tasks that focus on masked language modeling and multi-modal alignment prediction.
* Significant improvements are observable in downstream tasks, achieving state-of-the-art performance in visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval.
* The co-attentional transformer architecture allows for flexible interaction between visual and linguistic representations, accommodating their differing complexities.
* ViLBERT's architecture shows that a unified model can outperform traditional separate vision and language models, offering a more integrated understanding.
* Results demonstrate that the pretraining process enhances visiolinguistic representations, leading to better generalization in downstream tasks.
* The model's adaptability for various tasks requires minimal architectural changes, simplifying the process of transferring to new vision-and-language applications.

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 & Huang, Difeng. (2024). Image to Label to Answer: An Efficient Framework for Enhanced Clinical Applications in Medical Visual Question Answering. Electronics. 13. 2273. 10.3390/electronics13122273.
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. Multi-modal 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. J. Wu, F. Ge, P. Shu, L. Ma and Y. Hao, "Question-Driven Multiple Attention(DQMA) Model for Visual Question Answer," 2022 International Conference on Artificial Intelligence and Computer Information Technology (AICIT), Yichang, China, 2022, pp. 1-4, doi: 10.1109/AICIT55386.2022.9930294.
12. Z. Shao et al., "Visual Explanation for Open-Domain Question Answering With BERT," in IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 7, pp. 3779-3797, July 2024, doi: 10.1109/TVCG.2023.3243676.
13. S. Ravi, A. Chinchure, L. Sigal, R. Liao and V. Shwartz, "VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge," 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 1155-1165, doi: 10.1109/WACV56688.2023.00121.
14. D. Amin, S. Govilkar and S. Kulkarni, "Visual Question Answering System for Indian Regional Languages," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 22-27, doi: 10.1109/ICAST55766.2022.10039528.
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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No** | **Author Name** | **Paper Title** | **Date of publish-**  **ing** | **Observations** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Tan and Bansal | LXMERT: Learning Cross-Modality En- coder Representations from Transformers | 2019 | Introduced the LXMERT model, showing signif- icant improvements in VQA tasks by integrating visual and textual infor- mation. LXMERT uses three types of encoders: object-relationship, lan- guage, and cross-modality. Evaluated on datasets like VQA, GQA, and NLVR2, it outperformed prior models, showing superior perfor- mance in cross-modality tasks compared to previous BERT-like models focused  solely on language. |
| 2 | Liunian Harold Li et.al | VisualBERT: A Sim- ple Model for Vision- and-Language Tasks | 2019 | VisualBERT integrates BERT with image features extracted from object de- tectors, using Transformers to process text and im- age inputs. It undergoes task-agnostic pre-training, task-specific pre-training, and fine-tuning for spe- cific 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: Pretrain- ing Task-Agnostic Vision-and-Language BERT | 2019 | ViLBERT extends BERT with a two-stream ar- chitecture that processes textual and visual inputs separately and then inter- acts 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 align- ing 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 trans- former 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 An- swer: An Efficient Framework for En- hanced Clinical Appli- cations in Medical Vi- sual Question Answer- ing | 2024 | Med-VQA combines large language models and multi- label learning to address data scarcity and complex- ity in Medical VQA. The Image to Label to An- swer (ITLTA) framework reduces costs, improves in- terpretability, and enables zero-shot learning. Ex- periments on the VQA- Med 2019 dataset show that Med-VQA outperforms cur- rent approaches, enabling more effective clinical appli-  cations. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 7 | Rub’en Tito et. al | Hierarchical multi- modal transform- ers for Multi-Page DocVQA | 2023 | This paper extends Docu- ment VQA to multi-page documents with the MP- DocVQA dataset and pro- poses Hi-VT5, a hierar- chical multimodal trans- former. Hi-VT5 processes multi-page documents effec- tively and outperforms ex- isting models in terms of performance and explain- ability, making it a better fit for real-world applications involving multi-page docu-  ment contexts. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 8 | Aisha Urooj Khan et. al | MMFT-BERT: Multi-  modal Fusion Trans- former with BERT Encodings for Visual Question Answering | 2020 | In order to tackle the problem of Visual Ques- tion Answering (VQA), the MMFT-BERT model (Multimodal Fusion Trans- former with BERT en- codings) 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 per- forms better than earlier state-of-the-art models, demonstrating gains in  accuracy. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 9 | Wonjae Kim et.al | ViLT: Vision and Lan- guage 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 convo- lutional neural networks (CNNs) to streamline visual processing. ViLT embeds picture patches using a linear projection, in contrast to typical mod- els 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 complex- ity to make it up to ten times faster than previous  models. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 10 | Tim Siebert et. al | Multi-Modal Fu- sion Transformer for Visual Question Answering in Remote Sensing | 2022 | The paper presents a new architecture for visual ques- tion answering (VQA) in remote sensing (RS) called VBFusion, which jointly learns image and text rep- resentations by using multi- modal transformer models. By utilizing a feature ex- traction module, a fusion module that combines mul- tiple VisualBERT layers, and a classification module, this approach overcomes the limitations of existing mod- els and can handle complex, non-specific questions that go beyond predefined ob- ject categories. The results of the experiment show no- table 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 Mul- tiple Attention (DQMA) model is proposed in this paper. It deals with the problem of extraneous vi- sual information influenc- ing VQA models’ accuracy. The DQMA model uses LSTM for question features and Faster R-CNN for im- age feature extraction. Rel- evant image regions are cho- sen by a question-driven at- tention mechanism, which lowers noise. In order to improve the interac- tion between the question and image features, the model also incorporates co- attention networks. As- sessments conducted on the VQA 2.0 dataset reveal that the DQMA model surpasses alternative techniques, en-  hancing 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 An- swering (OpenQA) models, especially those that use BERT, the paper suggests VEQA, a visual analyt- ics system. OpenQA ad- dresses issues with compli- cated data and models while providing answers to queries derived from lengthy un- structured text passages. In order to address these, VEQA provides visual ex- planations at three different levels: summary, instance, and module. This en- ables experts to examine the way in which models handle questions and extract per- tinent passages. By using visual aids such as ranking visualizations and compara- tive trees, VEQA enhances the interpretability of mod- els and identifies areas that |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 13 | Sahithya Ravi et. al | VLC-BERT: Visual Question Answering with Contextual- ized Commonsense Knowledge | 2023 | The VLC-BERT model is designed for Visual Ques- tion Answering (VQA) with an emphasis on common- sense reasoning. It incorpo- rates contextualized knowl- edge using the COMET model, which is trained on human-curated knowl- edge bases like ConceptNet and ATOMIC. The model combines visual and textual inputs with commonsense inferences to answer ques- tions that require reason- ing beyond what is visible in the image. Through at- tention mechanisms, VLC- BERT selects the most rel- evant commonsense knowl- edge to enhance the VQA  task. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 14 | Dhiraj Amin et. al | Visual Question An- swering System for In- dian Regional Lan- guages | 2022 | The paper ”Visual Question Answering System for In- dian Regional Languages” addresses the scarcity of datasets for Indian lan- guages 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 trans- lated datasets and deep learning techniques can bridge the gap for regional  language VQA tasks. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 15 | Venkat Kodali et. al | Recent, Rapid Ad- vancement 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 devel- opments in Visual Question Answering (VQA). It high- lights how VQA, combining computer vision and nat- ural language processing, has grown exponentially in research output since 2015. The review covers key VQA models and techniques, including attention-based and transformer architec- tures, 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. |