

Visual Question Answering (VQA) – A Review

Harsh Bansal

(A25305221034)

B.Tech. Computer Science and Engineering (3rd Year),

Amity University Punjab,

Mohali, India

Email - harsh.bansal3@s.amity.edu

Abstract— This paper delves into the world of Visual Question Answering (VQA), exploring widely-used datasets that serve as the foundation for advancements in this domain. Seven distinct techniques, ranging from the renowned GPT-4 to the innovative BEiT-3, are comprehensively elucidated, showcasing the diversity and sophistication of contemporary approaches. These techniques represent the state-of-the-art methods in VQA, underscoring their prowess in addressing the nuanced interplay between vision and language. Amidst this progress, the abstract navigates the challenges inherent in VQA, addressing concerns such as dataset biases and the need for more precise results. By encapsulating discussions on datasets, cutting-edge techniques, and the intricate challenges faced, this paper offers a comprehensive overview of the current landscape in Visual Question Answering research.

Keywords—Visual Question Answering, VQA, Computer Vision, Multimodal AI, Object Detection, Natural Language Processing

I. INTRODUCTION

Visual Question Answering (VQA) is a captivating task that was introduced with the primary objective of forging a profound connection between computer vision and natural language processing (NLP). It serves as a catalyst for research, igniting innovation and pushing the boundaries of both fields. At its core, VQA seeks to bridge the gap between two distinct domains of artificial intelligence: computer vision, which focuses on methods for acquiring, processing, and comprehending visual data, effectively teaching machines how to 'see,' and NLP, which revolves around enabling human-computer interactions through natural language, imparting the ability to 'read' and understand human text. Although computer vision and NLP both reside within the broader realm of artificial intelligence and share common foundations rooted in machine learning, they have historically evolved in isolation, each dedicated to its unique set of challenges and objectives.[1]

Over the past few decades, both computer vision and NLP have achieved remarkable progress in their respective pursuits. Computer vision has made strides in developing methods to extract meaningful information from images, enabling machines to recognize objects, scenes, and patterns within visual data. Simultaneously, NLP has advanced in its mission to enable computers to understand, generate, and interact with human language, encompassing tasks as diverse as language translation, sentiment analysis, and text summarization. These parallel advancements have laid a solid foundation for the burgeoning field of VQA.

However, it is the confluence of these two domains, driven by the exponential growth of visual and textual data, that has given rise to a synergy between computer vision and NLP.

This convergence is exemplified by research in image captioning, where automatic image description has emerged as a potent technique. In this context, the approach revolves around the amalgamation of convolutional neural networks (CNNs), primarily trained for object recognition and feature extraction from images, with word embeddings, honed through extensive training on vast text corpora. This fusion of visual and textual information represents a significant stride in the realm of AI, enhancing the capacity of machines to comprehend and generate contextually relevant descriptions for visual content.

As we delve into the world of VQA in this paper, we will explore its evolution, underlying methodologies, and the synergy between computer vision and NLP that underpins its success. We will also delve into the challenges posed by VQA, its applications, and the potential it holds for reshaping the landscape of artificial intelligence by fostering a deeper connection between the domains of 'seeing' and 'reading.'[2].

II. AVAILABLE DATASETS

Visual Question Answering (VQA) has witnessed remarkable progress owing to the availability of diverse datasets that facilitate the training and evaluation of VQA systems. This section provides an overview of some prominent datasets in the VQA domain, highlighting their characteristics, sizes, and unique features.

A. DAQUAR

The Dataset for Question Answering on Real World Images (DAQUAR), introduced by Malinowski et al. [3], represents a pioneering effort in the VQA domain. Comprising 1449 images sourced from the NYU-Depth V2 dataset [11], DAQUAR features a question bank with 12468 question-answer pairs. These questions, derived from human annotations, are structured within 9 question templates, utilizing annotations from the NYU-Depth dataset.

B. VQA Dataset

The Visual Question Answering (VQA) dataset [4], one of the largest collections, draws images from the MS-COCO dataset [12]. With at least 3 questions per image and 10 answers per question, the dataset encompasses 614,163 questions in open-ended and multiple-choice formats. The multiple-choice questions categorize answers as Correct, Plausible, Popular, or Random. Additionally, VQA V2 introduces confusing images, expanding the dataset. Visual7W

Derived from the MS-COCO dataset, the Visual7W dataset [5] consists of 47,300 images and 327,939 question-answer pairs. Incorporating 1,311,756 multiple choice questions and answers with 561,459 groundings, Visual7W primarily

addresses seven question forms: What, Where, When, Who, Why, How, and Which. The dataset distinguishes between 'telling' and 'pointing' questions, enriching its diversity.

C. Tally-QA

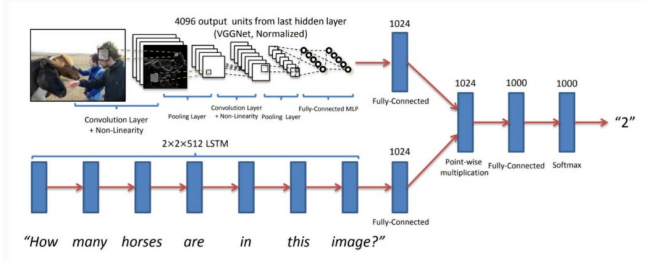
Introduced in 2019, the Tally-QA [6] dataset emerges as a substantial addition to VQA datasets, specifically addressing object counting in open-ended tasks. Boasting 287,907 questions, 165,000 images, and 19,000 complex questions, Tally-QA surpasses the VQA dataset in scale.

III. REVIEW OF TECHNIQUES

A. Vanilla Visual Question Answering (VQA) Approach

The Vanilla Visual Question Answering (VQA) approach, outlined in this section, serves as a benchmark for deep learning methodologies within the VQA domain. This model employs Convolutional Neural Networks (CNN) for feature extraction and utilizes Long Short-Term Memory (LSTM) or Recurrent Networks for language processing. The subsequent combination of these features through element-wise operations yields a unified feature representation, subsequently employed in the classification of answers. This section provides an in-depth exploration of the Vanilla VQA model, shedding light on its architecture and operational flow. For clarity, Figure 1 visually illustrates the operational flow of the Vanilla VQA model, showcasing the integration of CNN, LSTM or Recurrent Networks, feature fusion, and the ultimate answer classification process.

Fig. 1. Vanilla VQA



B. Stacked Attention Networks in Visual Question Answering

The Stacked Attention Networks, as proposed by [7], revolutionize the landscape of Visual Question Answering (VQA) by introducing a novel attention mechanism. This model leverages the softmax output of the intermediate question feature to establish attention, creating a dynamic focus on crucial elements within the image. The innovation lies in the stacking of attention mechanisms, enabling the model to discern and emphasize multiple significant portions of the visual input. This approach represents a paradigm shift in VQA, as it not only refines the interpretability of the model but also enhances its capability to understand and respond to intricate relationships within the visual and textual components of the input. The utilization of stacked attention networks reflects a nuanced understanding of image-question pairs, contributing to improved accuracy and robustness in answering a diverse array of questions. Refer Fig 2.

C. Teney et al. Model: Integrating Object Detection in VQA

The contribution of Teney et al. [8] to Visual Question Answering (VQA) is marked by a pioneering utilization of

object detection techniques, a paradigm that led them to victory in the VQA Challenge 2017. The model, depicted in Figure 3, introduces the incorporation of object detection within VQA architectures. By leveraging the R-CNN (Region-based Convolutional Neural Network) architecture, this approach strategically narrows down the features, enabling a more precise focus on pertinent elements within images. The use of object detection serves as a powerful mechanism to enhance attention, allowing the model to discern and prioritize relevant visual cues. Noteworthy is the demonstrated superiority of this model in terms of accuracy when compared to alternative architectures, underscoring the efficacy of integrating object detection for refined performance in the complex task of Visual Question Answering. This section delves into the architectural intricacies of the Teney et al. Model, shedding light on its pivotal role in achieving state-of-the-art results in VQA challenges.

Fig. 2. Stacked Attention Networks (SAN)

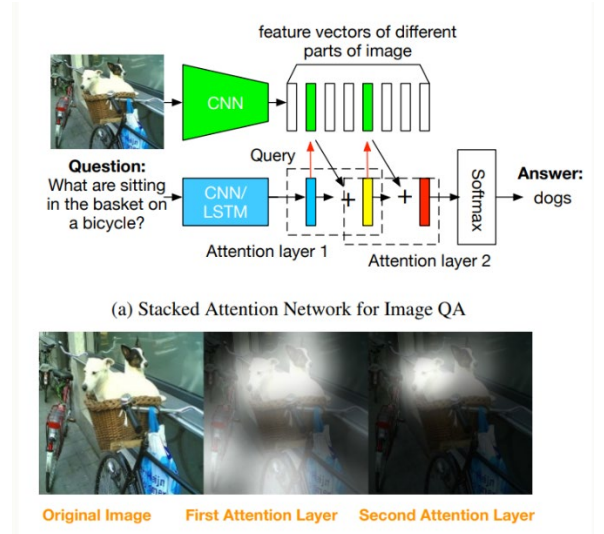
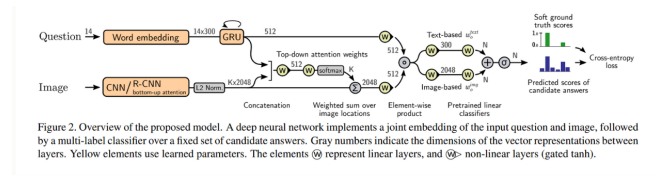


Fig. 3. R-CNN Architecture

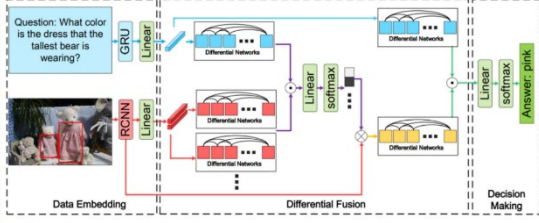


D. Differential Networks: Leveraging Differential Propagation for Enhanced Feature Learning

Differential Networks, as introduced by [9], represent an innovative approach in Visual Question Answering (VQA) that harnesses the differences between forward propagation steps to enhance feature learning. The model strategically employs Faster R-CNN for extracting image features, emphasizing precision in representation. Notably, differential modules are integrated into both textual and visual pathways, contributing to noise reduction, and fostering a deeper understanding of interdependencies between features. For question feature extraction, the model employs Gated Recurrent Unit (GRU), which effectively captures sequential dependencies within the textual input. The culmination of these features involves the integration of an attention module, culminating in the classification of answers. Figure 4 visually

encapsulates the architecture of Differential Networks, highlighting its unique approach to feature refinement and learning. This section provides an in-depth exploration of the model's architecture, elucidating how the interplay between differential propagation, feature refinement, and attention mechanisms collectively contribute to the model's efficacy in the challenging domain of VQA.

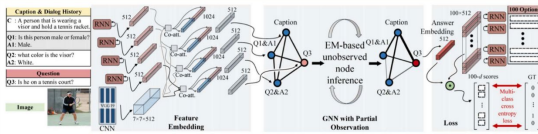
Fig. 4. Differential Networks Architecture



E. Differentiable Graph Neural Network (GNN) for Modeling Visual Dialogs

A recent advancement in Visual Question Answering (VQA), presented by Zheng et al. [10], introduces the Differentiable Graph Neural Network (GNN) as a novel framework for modeling visual dialogs. This innovative approach conceptualizes visual dialogs as structural graphs and Markov Random Fields. In this framework, the entities involved in the dialog are represented as observed nodes, and the answer assumes the role of a node with a missing value. The Differentiable Graph Neural Network (GNN) architecture, depicted in Figure 5, showcases the intricate interplay of these components. By structuring visual dialogs in this manner, the model aims to capture and leverage the inherent relationships and dependencies within the dialog, paving the way for a more nuanced understanding and improved accuracy in answering questions. This section delves into the specifics of the Differentiable GNN model, elucidating its unique approach to representing visual dialogs and its potential implications for advancing the state-of-the-art in VQA.

Fig. 5. Example of a figure caption. (figure caption)



IV. LATEST ADVANCEMENTS IN VQA

A. GPT-4 by OpenAI

GPT-4 represents a significant leap in the realm of language models by extending its capabilities to multimodal inputs—encompassing both images and text. As shown in Figure 6, This groundbreaking feature allows users to articulate prompts that involve a seamless integration of visual and textual elements, empowering the model to tackle diverse vision and language tasks. Irrespective of the complexity, GPT-4 exhibits prowess in generating text outputs when presented with inputs consisting of intricately interleaved text and images. The model's performance spans various domains, encompassing documents that combine textual content with photographs, diagrams, or screenshots. Table 3 provides a glimpse of GPT-4's adept handling of

visual inputs. Notably, the standard test-time techniques, such as few-shot prompting and chain-of-thought, established for language models, prove equally effective in the multimodal setting.

Fig. 6. Example of GPT-4 Visual Input

Example of GPT-4 visual input:

User What is funny about this image? Describe it panel by panel.



Source: <https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/>

- GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.
- Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.
- Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.
- Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.
- The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

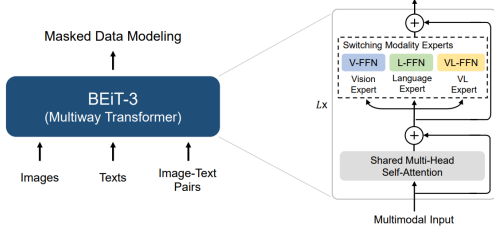
Preliminary assessments of GPT-4's visual capabilities extend to a narrow set of academic vision benchmarks, with initial insights available in the GPT-4 blog post. Anticipating the growing interest in this groundbreaking fusion of vision and language, subsequent works are planned to unveil more comprehensive details about GPT-4's performance, shedding light on its potential applications and impact across diverse domains.[13]

B. BEiT Pretraining for All Vision and Vision-Language Tasks

BEiT-3 stands at the forefront of multimodal pretraining, presenting a state-of-the-art solution for both vision and vision-language tasks. Crafted by a team at Microsoft Corporation, BEiT-3 introduces innovations in three key dimensions: backbone architecture, pretraining task, and model scaling. Utilizing Multiway Transformers, the model achieves versatile and effective general-purpose modeling through a modular architecture supporting deep fusion and modality-specific encoding. The distinctive feature lies in its unified pretraining strategy, employing masked "language" modeling on images (Imglish), texts (English), and image-text pairs ("parallel sentences"). This comprehensive approach propels BEiT-3 to excel across diverse benchmarks, demonstrating superior performance in object detection, semantic segmentation, image classification, visual reasoning, visual question answering, image captioning, and cross-modal retrieval. The model's prowess positions it as a

cutting-edge solution, pushing the boundaries of multimodal AI applications. For in-depth exploration, the official publication is available at BEiT-3.[14]

Fig. 7. Overview of BEiT-3 Pretraining



V. EVALUATION METRICS

Evaluation metrics play a pivotal role in assessing the performance of models across various tasks. In the context of classification and information retrieval, several key metrics are commonly employed, each providing unique insights into different aspects of model performance. The following metrics—Accuracy, Precision, Recall, and F1-Score—are fundamental in quantifying the efficacy of models in scenarios ranging from image classification to natural language processing.

A. Accuracy

Accuracy represents the ratio of correctly predicted instances to the total instances. It is a comprehensive metric that measures the overall correctness of the model. The formula for accuracy is given by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

B. Precision

Precision assesses the accuracy of positive predictions made by the model. It is calculated as

$$Precision = \frac{TP}{TP + FP}$$

where TP is the number of true positives and FP is the number of false positives. Precision is particularly useful in scenarios where false positives need to be minimized, such as medical diagnoses or spam detection.

C. Recall

Recall, also known as sensitivity or true positive rate, measures the ability of the model to capture all relevant instances. It is calculated as

$$Recall = \frac{TP}{TP + FN}$$

where TP is true positives and FN is false negatives. Recall is crucial in situations where missing positive instances is costly, such as identifying rare diseases or security threats.

D. F1-Score

F1-Score is the harmonic mean of precision and recall, providing a balanced assessment of a model's performance. It is calculated as

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

F1-Score is particularly valuable when there is an imbalance between positive and negative instances, offering a consolidated measure that considers both precision and recall.

VI. CURRENT CHALLENGES AND RESEARCH OPPORTUNITIES

Visual Question Answering (VQA) stands apart from traditional computer vision tasks like object detection, where the focus is on classifying objects in an image without addressing natural language queries related to the image content. Current state-of-the-art VQA systems assume question relevance to the image and often confidently provide incorrect answers to irrelevant questions. Addressing this, VQA systems should prioritize checking the relevance of the posed question before responding. However, existing VQA techniques predominantly emphasize image processing tasks, with limited attention to understanding the language semantics of the questions. Research gaps in this area include the need for domain-specific datasets, challenges in salient object detection, the absence of semantic relationships among image objects leading to uncertainty in relevant answers, and the necessity for joint learning of VQA and Question Relevance Detection (QRD). Additionally, concerns arise from biased datasets, the overconfidence of present VQA tasks in answering irrelevant questions, and the demand for more precise and accurate results in VQA. These gaps motivate further research and exploration in the field.[2]

VII. CONCLUSION

In conclusion, Visual Question Answering (VQA) emerges as a distinct paradigm in computer vision, diverging from traditional tasks such as object detection by addressing natural language queries related to image content. However, the current state-of-the-art VQA systems exhibit limitations, confidently providing incorrect answers to irrelevant questions, thereby emphasizing the need for relevance checking before responses. Notably, existing VQA techniques predominantly focus on image processing tasks, with less emphasis on understanding the nuanced semantics of posed questions. Research gaps in the field include the demand for domain-specific datasets, challenges in salient object detection, and the absence of semantic relationships among image objects leading to uncertainty in relevant answers. The call for joint learning of VQA and Question Relevance Detection (QRD) underscores the need for more comprehensive approaches. Additionally, concerns about biased datasets, overconfidence in answering irrelevant questions, and the pursuit of more precise and accurate results further propel the need for continued exploration and research in the dynamic realm of VQA. As we navigate these challenges and opportunities, the evolution of VQA promises

to reshape the landscape of multimodal AI applications, fostering advancements that extend beyond traditional computer vision boundaries.

REFERENCES

- [1] Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2019). Visual Question Answering using deep learning: A survey and performance analysis. doi:10.48550/ARXIV.1909.01860
- [2] Chirgaiya, S., Rajavat, A. (2021). Comprehensive Analysis of State-of-the-Art Techniques for VQA. In: Kiran Mai, C., Kiranmayee, B.V., Favorskaya, M.N., Chandra Satapathy, S., Raju, K.S. (eds) Proceedings of International Conference on Advances in Computer Engineering and Communication Systems. Learning and Analytics in Intelligent Systems, vol 20. Springer, Singapore. https://doi.org/10.1007/978-981-15-9293-5_9
- [3] Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: NIPS, pp. 1682- 1690 (2014)
- [4] Antol, S. , et al.: VQA: visual question answering. In: IEEE ICCV, pp. 2425- 2433 (2015)
- [5] Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: grounded question answering in images. In: IEEE CVPR, pp. 4995-5004 (2016)
- [6] Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. arXiv preprint arXiv:1810.12440 (2018)
- [7] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: IEEE CVPR, pp. 21-29 (2016)
- [8] Terrey, D., Anderson, P., He, X., van den Hengel, A.: Tips and tricks for visual question answering: learnings from the 2017 challenge. In: IEEE CVPR, pp. 4223- 4232 (2018)
- [9] Wu, C., Liu, J., Wang, X., Li, R.: Differential networks for visual question answer□ing. In: AAAI 2019 (2019)
- [10] Zheng, Z., Wang, W., Qi, S., Zhu, S.C.: Reasoning visual dialogs with structural and partial observations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6669---6678 (2019)
- [11] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 746- 760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
- [12] Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740-755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
- [13] OpenAI. (2023). GPT-4 Technical Report (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2303.08774>
- [14] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., & Wei, F. (2022). Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2208.10442>
- [15] Manmadhan, S., & Kovoov, B. C. (2020). Visual question answering: a state-of-the-art review. In Artificial Intelligence Review (Vol. 53, Issue 8, pp. 5705–5745). Springer Science and Business Media LLC. <https://doi.org/10.1007/s10462-020-09832-7>