THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 5. Issue 1.1 (2025)
https://doi.org/10.62019/1h3r5m91

Analysis of Deep Learning Models on VQA(Toloka) Dataset

Fariha Shoukat*, Muhammad Naveed, Nazia Azim, Arooj Imtiaz

Chronicle

Article history
Received: July 24, 2025
Received in the revised format: Aug 25, 2025
Accepted: Sept 13, 2025
Available online: Oct 6, 2025

Fariha Shoukat & Arooj Imtiaz are currently affiliated with the Department of Data Science, Riphah Institute of Systems Engineering, Riphah International University, Islamabad, Pakistan.
Email: farihashoukat421@gmail.com
Email: aroojkhan001@yahoo.com

Muhammad Naveed is currently affiliated with the Department of Computer Science, Virtual University, Pakistan, Pakistan.
Email: naveedalm2@gmail.com

Nazia Azim is currently affiliated with the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan.
Email: n.azim@awkum.edu.pk

Corresponding Author*

Abstract

Visual Question Answering (VQA) is a recent task at the intersection of computer vision and natural language processing. In this task, instead of giving phrase images, machines will go on to learn and interpret them. It is also their job to answer any question in turn. Instead of the usual "Images ten questions", it relies on phrases reversed in order and asks unrelated questions about pictures. One way which VQA current challenges itself picks up new trends is with this issue for the model to predict an image and where provide bounding box as output, given input image together with a questioning on natural language sentence. Unlike traditional VQA tasks, where the model outputs a textual answer, this challenge presents a more complex scenario by not explicitly mentioning the object name in the question, making it more demanding than typical VQA and visual grounding tasks. A comparative analysis is performed using the Toloka VQA challenge dataset, which has seen limited exploration in existing research. Our findings show that a combination of the DINO and T5 models achieved a high Intersection over Union (IoU) score of 0.67, demonstrating its effectiveness in this unique context. Similarly, the BERT combined with Swin Transformer model yielded a competitive IoU score of 0.60. In contrast, models like MobileNet-v2 and GPT with ResNet-50 + BERT struggled, displaying lower IoU values and highlighting the challenges posed by this dataset.

Keywords: Toloka VQA, IoU, Visual Question Answering, Object detection.

INTRODUCTION

When it comes to artificial intelligence, computer vision combined with natural language processing is now changing all kinds of application. Visual question answering is an interdisciplinary field that seeks to imbue machines with the ability to understand images and answer interrogative textual queries concerning them.VQA holds great promise, with applications ranging from flexible tools for the visually impaired to intelligent image recognition systems (Carion et al., 2020). But its full potential is dampened by intrinsic challenges. Generally speaking, Visual Question Answering combines images and text. Visual question answering systems are designed to give precise, natural-language responses to queries about both an input image and accompanying text. The main idea behind the problem is to design automatic system that can understand the contents of visual image and question as humans do (Ding et al., 2022) . It is very easy for humans to understand image content but difficult for machines. VQA combines both vision and natural language input data and requires understanding and correct interpretation of both using various deep learning models (Dosovitskiy et al., 2020). In this research, the focus is on both VQA and object detection because VQA is a challenge in which for each image, a natural language question is given, and the task is to examine image features and draw a bounding box around the image which is to answer the question. The evaluation of

these types of tasks is done using Intersection over Union. There is a huge gap between human prediction and deep learning model prediction, and it is interesting to see especially after huge advances in multi- modal deep learning models that involve computer vision language in the last few years. A comparative analysis between various advanced deep learning model performances is performed to reduce the gap between human and deep learning models prediction (Du et al., 2022).

An important concept in deep learning is object detection at all scales. Identifying and annotating objects in photos and videos in the process of object detection means that each recognized kind of object carries with it some other bunch on top of itself-- a category tag. Object detection for visual question answering (VQA) means identifying and locating objects in an image so that accurate answers can be generated to questions about the visual content. Visual question answering is challenging because it involves understanding not only the text query but also an image's visual content. Combining object detection techniques with VQA models allows them to understand the visual context better, providing more informed and accurate answers to questions about images. It even assists in providing a hierarchical decomposition of visual elements in an image. When objects and their locations are identified, the VQA model is better able to know other things about the scene as well, such as which part corresponds to which object type.

If whether an object exists in both visuals and text gives rise to doubt, then this helps clarify things because having seen that certain items are alike simply functions as a rectification (Deng et al., 2021). In this research, we used WSDM2023 Toloka VQA challenge dataset which is very different from other datasets because of its interrogative nature. In other words, the category name of objects doesn't explicitly appear in questions, which makes it more challenging than common visual grounding tasks (Gómez Blanco et al., 2024). It is a challenge in which for each image, a natural language question is given, and our task is to examine image features and draw a bounding box around the object in the image, which is to answer the question. The evaluation of these types of tasks is done using Intersection over Union. On Toloka VQA dataset, the human performance was 0.87 while deep learnings models scores were only 0.21. So, there was huge accuracy gap between human and deep learning models performance (Kamath et al., 2021).

For comparative analysis between various advanced deep learning models. Four different combinations of text and image-based models such as DINO with T5, Swin Transformer and BERT, ResNet-50 with BERT, and MobileNet-v2 with GPT are selected. Initially the model is trained on 21927 and 22667 training samples and validated models on 8323 samples. In our case, we achieved the highest IoU score of 0.6777 on validation set using DINO and T5. The second most successful combination was Swin Transformer and BERT for which we achieved IoU score of 0.60. The other two combinations, ResNet-50 with BERT and MobileNet-v2 with GPT did not perform well for this problem.

RELATED WORK

A. Deep Learning Techniques for VQA

VQA with object detection has a variety of techniques used. The method of Object Detection has been proposed in (Lin, Chen, Mei, Coca, & Byrne, 2024) as a radically new object detection technique. Historically, classifiers were used in object detection tasks. However, object detection has been treated as a regression problem by these

approaches. The focus is on predicting spatially separated bounding boxes and their class probabilities. In one fell swoop, a neural network localizes the objects of interest and outputs class probabilities from scratch on entire images. After a couple of starts and stops, detection performance may be accommodated since the whole detection pipeline is a single network. A feature-wise attention mechanism in VQA was newly introduced in this paper, with more attention paid to relevant characteristics, less attention to irrelevant ones, and better discriminative features for image and question both.

Accordingly, we invented new modules of VQA on this stage as well as off-shelf this larger coattail network: the Union Feature-wise and Spatial Co-Attention Network. Their VQA model can achieve strong fine-grained recognition with attention at different dimensions. Experiments on two large-scale real-world VQA datasets show that their approach has surpassed the state of the art (Lu, Ding, Liu, Yin, Yin, & Zheng, 2023) . Attention-based Models, such as the Bottom Up and Top-Down Attention model (Liu et al., 2021), have become common in VQA. These models use attention mechanisms to weigh the importance of different regions in an image when answering a question. In our work, a method for implementing attention mechanisms from both bottom-up and top-down angles is proposed. This is a way to calculate attention at object level and other key positions in the picture In (Lin, Zhang, Tao, Shi, Haffari, Wu, He, & Ge, 2023), a novel mobile architecture was presented that improves the performance of advanced mobile models across a range of model sizes, tasks, and benchmarks.

Additionally, discussed effective strategies for utilizing these mobile models for object identification in a cutting-edge framework named SSDLite. Moreover, a demonstration is done on how to generate mobile semantic segmentation models using a trimmed-down version of DeepLabv3. The research study in (Lu, Batra, Parikh, & Lee, 2019) showed that image classification tasks can be successfully accomplished without CNNs when a pure transformer is applied directly to picture patch sequences. In (Lu, Liu, Yin, Yin, Liu, & Zheng, 2023) , the authors identified these challenges as three distinct attention problems and proposed an Accumulated Attention (A-ATT) mechanism to collectively address them. Their A- ATT technique may gradually disregard disturbances while continuously accumulating attention for relevant information in images, queries, and objects.

Models based on the Transformer architecture, such as the ViLBERT model (Pei et al., 2020) , have shown excellent performance in VQA. As a result, these models are particularly well-suited to tasks that require fine-grained attention and language modeling. Graph Convolutional Networks: Image objects were linked by relation and VQA model performance was promoted in multiple cases through graph convolutional networks (GCNs). The performance of VQA models with object detection can be improved using GCNs, as demonstrated by (Qian et al., 2024) (Schwenk et al., 2022) employs a two-stage strategy with segmentation after detection. The authors separately trained a more robust segmentor and detector. Additionally, they implemented pseudo-label training on the test set using a student-teacher framework and utilized object detection based on end-to-end transformers. A cutting-edge end-to-end object detector DINO, was introduced by (Shao et al., 2023) . DINO took the look ahead twice technique for box prediction into use, so it achieved both better performance and higher efficiency than previous DETR-like models. DINO also used a mixed query selection method for anchor initialization and a contrastive strategy to improve overall training accuracy. But just like the window method

announced after departure of a standard (see footnote 5), DETR improves performance and yet removes all of its largely hand-designed components that need bug-prone tuning before it can be deployed successfully--while giving you something else for a change. But due to the fact that Transformer attention modules cannot analyze picture feature maps, it becomes slow to converge and has low-feature spatial resolution.

To address these problems a solution of attentions calculated more locally than by necessity was proposed in (Ustalov et al., 2023) called Deformable DETR, in which its attention modules only focus on a limited number on small sample points around one tollbooth out of many different spots nearby (figure 9). When dealing with small objects, Deformable DETR gets better results than DETR by 10% less training epochs. In reference (Wang et al., 2021) proposes MDETR, a top-down modulated object detector. An example of its operation is shown below. It takes a raw text query such as `search for the item behind the rod to the left'' or `What is in this picture?'' came in from the user and locates objects in the situation figured by this natural language description of which parts and at what locations they are supposed to occur. It combined the two modalities at an early point in the model (so off-sets could be given where needed), and allowed for text and image to be jointly processed by a transformer-based design.

Developed a revolutionary method for handling the detection task. In their paper (Zhang, Zhang, & Xu, 2023). object recognition is treated as a direct set prediction problem By getting rid of many bibliography related artifacts the authors' results not only are theoretically significant; but their approach also has practical consequences. The authors provide an algorithm with improved complexity for the detection pipeline based on integrated end-to---end convolutional neural nets and faster R-CNN The Pyramid Vision Transformer (PVT) was introduced in (Zhang, Chen, Chen, Zou, Li, & Lu, 2021).

PVT overcomes a number of challenges faced when transferring the transformer to all kinds of dense prediction systems. Unlike conventional hybrid approaches or pure transformer ones it supplies reasonable workarounds to these problems. PVT has several advantages over the state-of-the-art at present. PVT, not like ViT, goes to the trouble to perform the necessary number of operations in largely feature maps - opening up this part of an image at high resolution via dense 15 splitting and rapid computation. ViT usually requires much memory and a lot of calculations to give poor performance at a low output resolution. PVT combines the strong features of Transformer and CNN to form a more comprehensive framework for vision work that can be done without convolutions.

Introduced in paper 22, BERT is a new language representation paradigm. All layers of training BERT at the same time basically means left and right context inputs: this is intended to pre-train very deep contextual representations in contrast with the recent models only with situations. Therefore, for BERT in applications that add pre-training tasks specific to given domains but do not change its architecture, the model can be quite easily fine-tuned with just one additional output layer and then achieves state-of-the-art results on a wide array of test suites, including question answering and natural language inference cooperation between two or more different entities. In paper 23, the authors propose a transformer-based visual method. Their approach feature the transformer encoder-encoder, and did not use pre-trained detectors or word embedding models like those proposed in existing proposal-and-rank frameworks based on pre- trained object detectors in proposal-free frameworks

improving off the shelf one stage detector by incorporating textual embeddings. Using Multimodal Compact Bilinear Pooling (MCB), paper 24, the authors combine the two types of modalities. In the architecture proposed for Visual Question Answering(VQA), which consists of bundles of Multimodal Compact Bilinear pooling (MCB) units and attention mechanisms, we have superiority compared with all other existing methods in VQA datasets. Introduction of MCB pooling improves phrase localization accuracy in visual grounding tests, indicating the existence of steepened interaction between query phrase representations and visual representations for proposals 'bounding boxes.

MATERIAL & METHODS

Multiple steps are performed to achieve the proposed solution.

Data Pre-processing and Feature extraction

In pre-processing, questions are tokenized into words, and images are also preprocessed such as changes in shape, size, and cropping. Tokenization is employed in natural language processing to split the sentences and the paragraphs into smaller parts that are easier to give meaning. The first step in the natural language processing (NLP) process is to break down a given piece of text into an understandable format by dividing the in formativeness of the sentence into intelligible units, such as words. This division helps in making the text more comprehensible for further processing in NLP. The goal of pre-processing of an image is the enhancement of the image data which reduces unwanted distortions or amplifies some image features that may be useful for subsequent stages and analysis task. First of all, the dataset is downloaded from Toloka official website. The dataset has images URLs only, so we need to download images from the CSV files before further processing. Due to large images data size, it was very challenging to download images. Our dataset comprises various columns, including image URLs, questions, width, height, and bounding box coordinates (top, left, right, and bottom).

Table 1.
Overview of Toloka VQA Dataset

Column Names	Data Type	Description
Image	String	Images link
Question	String	Questions in English language
Width	Integer	Width of image
Height	Integer	Height of image
Left	Integer	Left coordinate of bounding box
Top	Integer	Top bounding box coordinate
Bottom	Integer	Bottom coordinateof bounding box
Right	Integer	Right coordinate of bounding box

After data downloading, the pre-processing is performed. To handle missing images in the image directory, we first identified the rows having missing images, created a new data frame and the image exists. In the image data frame, we categorized images into two groups: "true" and "false," based on whether the images existed. To prevent errors during model training, we calculated the counts of missing images and removed those rows from the dataset before feeding the data into the models. Additionally, we addressed any corrupted images using exception handling. Images were resized according to the specific requirements of each model and converted into tensors. We also generated word clouds to represent the most frequently used

words in the questions and visualized the character counts.

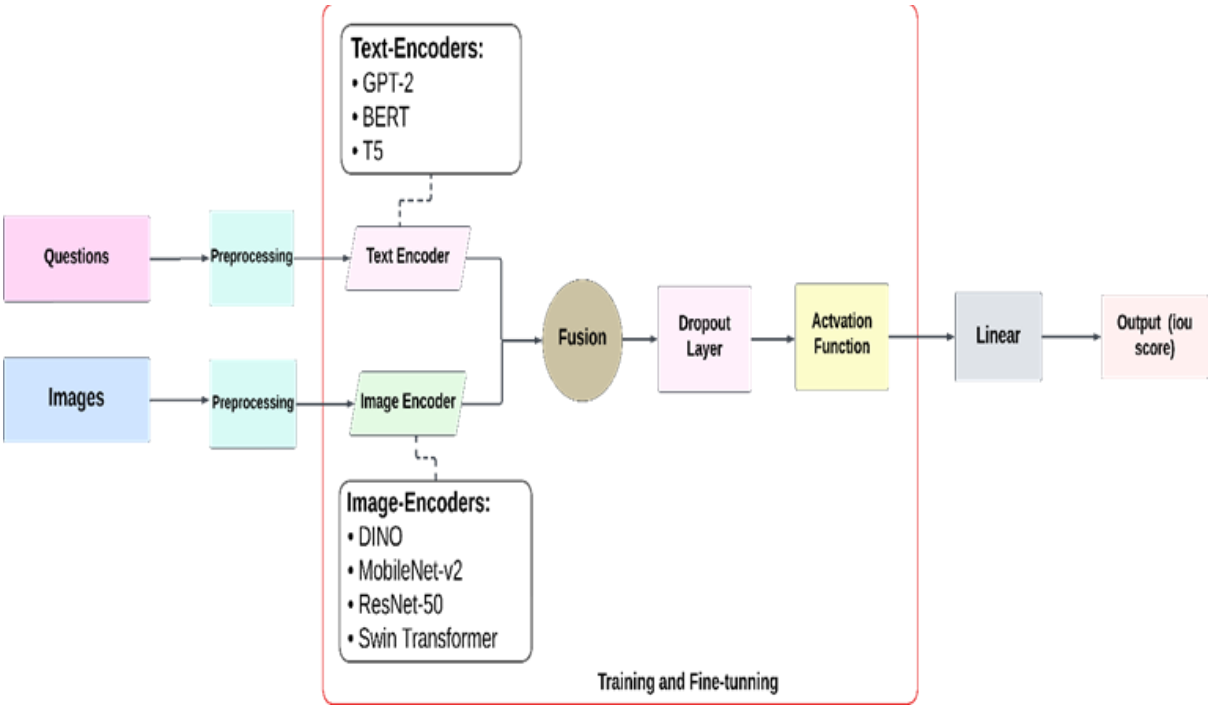

Figure 1.
This figure presents the overview of proposed methodology and steps to solve Toloka

Exploratory Data Analysis

In this step, we conducted an exploratory analysis of the dataset using various functions. We examined the dataset's shape and size, as well as performed word count and character count analyses. Additionally, we determined the maximum character counts for the questions in the dataset to gain a deeper understanding of the data's structure and content.

Visualization

We plotted character counts for questions that we calculated in previous steps. X-axis represents questions

while y-axis represents character counts. Created word clouds for the most common words used in questions. We can see that 'use' and 'used' are the most common words. We can also see that the data questions contain mostly verbs instead of object names.

Model Choice and Implementation

We have two types of models to deal with images and questions hence we tried different combinations of models such as DINO with T5, Swin Transformer with BERT, ResNet-50 with BERT and MobileNet-v2 with GPT-2.0.

Image-based Models

1. Self-Distillation with No labels (DINO)

The DINO model is self-supervised. It trains deep neural networks without needing any labeled data. DINO is built upon the architecture of Vision Transformers (ViT), utilizing the Transformers backbone in image understanding. DINO maintains within its

characterization a unique essence that is data augmentation, cluster assignment and contrastive loss. Images are often distorted at random when we input them for training. Which brings in more of the same picture. This means a lot less information changes between each pair of views made from one original photo. Your various transformations make it easier for the model to identify that which is worth keeping invariant to process any further. Images that resemble one another closely are placed in clusters.

Swin Transformer

In the field of computer vision, Swin Transformer is a unique architecture that is especially intended for image identification applications. Here are the details of Swin transformer.

- **Hierarchical Processing:** The input image is divided into multiple-scale, non-overlapping patches using Swin Transformer. Through a hierarchical processing mechanism, these patches enable the model to efficiently capture both local and global data.

- **Shifted Windows:** Swin Transformer uses shifted windows as opposed to standard transformers, which use fixed-size square patches. By moving the windows around in the image, these windows let the model better utilize the available data by capturing spatial correlations.

- **Step-by-Step Processing:** The model is divided into multiple stages, each of which has a number of transformer layers. The patches are processed independently inside each step, and cross-stage links are created to facilitate communication across various resolutions or scales.

- **Transformers with tokenization:** By tokenizing the picture patches into sequences, the Swin Transformer can handle the image as a collection of tokens, just like it would in tasks involving natural language processing. These tokens are processed by transformer layers, which capture intricate interactions between image parts.

- **Efficiency and Scalability:** Swin Transformer is computationally efficient and scalable to handle higher input resolutions, allowing it to achieve outstanding performance on image recognition applications.

- **Performance:** On a number of benchmark datasets, the architecture has shown competitive performance against transformer-based models and convolutional neural networks (CNNs), demonstrating its capacity to efficiently capture spatial dependencies and produce state-of-the-art outcomes in image recognition tasks.

Text-based Models

1) Text-to-Text Transfer Transformer (T5)

Based on the Transformer architecture, T5 is a flexible and potent language processing model that incorporates feed- forward neural networks and a self attention mechanism. Transformers are very effective for problems involving sequences because they enable the model to process input data in parallel. It also includes a multi-layered encoder and decoder. After processing incoming text, the encoder converts it to a fixed size representation. Using the task prefix and encoded text as input, the decoder produces output based on that fixed-sized representation.

2) Bi-directional Encoder Decoder Transformer (BERT)

Intending to pre-train deep bidirectional start from the raw text, and the desire is that every layer uses both left context every time as well as of course previous (right) outputs. Therefore, we can conclude that by simply appending another output layer to the pre-trained BERT model, discussions and answers to questions of all kinds have until now usually been state of the art.':

Models Training and Fine-Tuning

The concatenation method is employed to fuse features obtained from the models and trained models to calculate IoU score. For models training and fine-tuning, we used different numbers of epochs (10, 20, 30), learning rates (1e- 6, 1e-5), batch sizes (8, 16, 32) and training data samples. We trained the models on different numbers of training samples and fine-tuned using different hyper-parameters. Initially we trained models on small training samples. DINO with T5 is trained on 21927 and 22667 training samples and Swin Transformer with BERT is trained on 21927 training samples. For validation, 8323 samples are used.

Table 2.
Hyper parameters for models fine-Tuning

Batch Size	Learning Rate	Number	of Epochs	Optimizer
8	1e-5	20/30		Adam-W
16	1e-5, 1e-6	20/30		Adam-W
32	1e-5	10/20/30		Adam-W

- Batch Size: This is the number of samples input into the model before it takes a seach. Smaller batch sizes can lead to more generalization, while larger batch sizes can speed up training but require more memory.

- Learning Rate: It's agreed to regulate matters in which steps it takes our optimization method. With a perfect learning rate one can ensure that there will be no unnecessary oscillation around an optimal point or slow convergence. It helps models avoid over-fitting problem.

- Number of Epochs: This shows how many signify the logical sequel once these patterns are done emerging from the dataset (online). It affects a model's ability to recognize such structures within the dataset.

Figure 2.
Training IoU score obtained from Swin Transformer and BERT

Optimizer Weight decay is built directly into the optimization process of Adam by Adam- W, an extension of the Adam optimizer. Adam-W applies weight decay (L2 regularization) straight to the optimization process, hence addressing the problem of weight decay in Adam. By punishing excessive weights, this helps to prevent over

fitting and encourages a more robust model.

Evaluation and Performance Metric

Performance evaluation metrics are those metrics which help in measuring and evaluating the performance of a model. There are a lot of performance metrics available for evaluating the models. Choosing the best performance metrics based on the problem is one of the most important and challenging steps. Intersection over Union Score is the selected performance metric for evaluation.

IOU = Area of Intersection Area of Union

RESULT & DISCUSSION

For all models, the evaluation metric that was used is Intersection over Union (IoU). IoU is used mainly in applications such as object detection where in the actual model is trained to produce a frame that would surround an item in a proper manner.

Table 3.
Comparative Analysis among the Models based on IoU score on data samples

Models Applied	Validation IoU
DINO and T5	0.6777 (10 epochs, batch size 32, learning rate = 1e-6)
	0.49 (30 epochs, batch size 32, learning rate = 1e-5)
Swin Transformer and BERT	0.60 (20 epochs and 16 batch size, learning rate = 1e - 5)
	0.58 (30 epochs)

With our research, we hoped to advance the study of Visual Question Answering. The Toloka VQA dataset was very different from ordinary VQA dataset because of its interrogative nature. Our main objective was to carry out comprehensive comparison studies and resolve the accuracy gap issues between human and AI models. It was quite challenging because even after huge advancement in AI, models were not performing well on this dataset. The following plot shows the training IoU scores for Swin and BERT models where x-axis shows the number of epochs used to train models and y-axis shows IoU scores. We applied a combination of different models including the most advanced models such as Swin Transformers to solve this problem and succeed in reducing the accuracy gap and improving IoU scores.

Using this dataset to close the performance gap between humans and machines is the major step toward the development of AI systems that can comprehend and react to interrogative queries in visual situations. After performing multiples experiments using different model combinations, DINO with T5 outperformed on this dataset and achieved IoU scores of 0.67 on validation dataset. Below plot shows the validation IoU and loss scores. The numbers of epochs were 10 and training data samples were 22667. For validation, 1000 data samples are used. The x-axis represents the number of epochs and y-axis represents IoU score and loss value obtained during model evaluation.

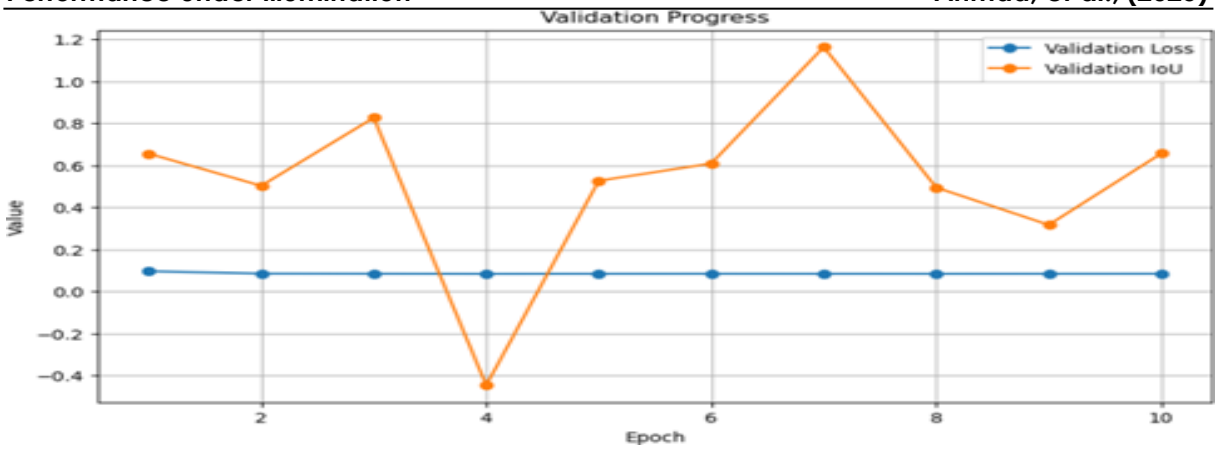

Figure 3.
Validation IoU score obtained using DINO and Text-to- Text Transfer

The second most successful model combination was Swin Transformer and BERT which are also very advanced models, achieved IoU scores of approximately 0.60 whereas the initial IoU scores were 0.21 using YOLOR and CLIP. Initially, we did not train our models on the complete dataset due to resource limitations.

CONCLUSION

The analysis of the Toloka VQA challenge dataset reveals the key findings: models like DINO + T5 achieved an IoU score of 0.67 across 21,000 data samples, demonstrating strong performance on this dataset. The BERT and Swin Transformer combination has comparatively better performance, with an IoU score of 0.60 across 22,000 samples. In contrast, MobileNet-v2 and GPT with ResNet- 50 + BERT showed lower IoU scores, indicating difficulties with the dataset's unique requirements. These results highlight the importance of selecting appropriate model architectures based on the dataset's characteristics. Future work could focus on optimizing these models or exploring alternative strategies to enhance performance on similar question-answering datasets.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.
Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.
Availability of data and material: In the approach, the data sources for the variables are stated.
Authors' contributions: Each author participated equally to the creation of this work.
Conflicts of Interests: The authors declare no conflict of interest.
Consent to Participate: Yes
Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213–229). Springer.

Deng, J., Yang, Z., Chen, T., Zhou, W., & Li, H. (2021). TransVG: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1769–1779).

Ding, Y., Yu, J., Liu, B., Hu, Y., Cui, M., & Wu, Q. (2022). MUKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5089–5098).

Du, Y., Fu, Z., Liu, Q., & Wang, Y. (2022). Visual grounding with transformers. *IEEE*, 1–6.

Gómez Blanco, R., Pérez Peinador, A., Sanjuan Espejo, A., Sánchez-Ruiz, A. A., & Díaz-Agudo, B. (2024). Experiential questioning for VQA. In *International Conference on Case-Based Reasoning* (pp. 305–320). Springer.

https://doi.org/10.1109/ICCV48922.2021.00982

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., & Carion, N. (2021). MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1780–1790).

Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2024). Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems, 36*.

Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., & Ge, Z. (2023). Medical visual question answering: A survey. *Artificial Intelligence in Medicine, 143*, 102611. https://doi.org/10.1016/j.artmed.2023.102611

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X., & Zheng, W. (2023). The multimodal fusion in visual question answering: A review of attention mechanisms. *PeerJ Computer Science, 9*, e1400. https://doi.org/10.7717/peerj-cs.1400

Pei, H., Wei, B., Chang, K. C.-C., Lei, Y., & Yang, B. (2020). Geom-GCN: Geometric graph convolutional networks. *arXiv preprint* arXiv:2002.05287.

Qian, T., Chen, J., Zhuo, L., Jiao, Y., & Jiang, Y.-G. (2024). nuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence, 38*, 4542–4550.

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., & Mottaghi, R. (2022). A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision* (pp. 146–162). Springer.

Shao, Z., Yu, Z., Wang, M., & Yu, J. (2023). Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14974–14983).

Ustalov, D., Pavlichenko, N., Likhobaba, D., & Smirnova, A. (2023). *WSDM Cup 2023 challenge on visual question answering*.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 568–578).

Zhang, S., Chen, M., Chen, J., Zou, F., Li, Y.-F., & Lu, P. (2021). Multimodal feature-wise co-attention method for visual question answering. *Information Fusion, 73*, 1–10. https://doi.org/10.1016/j.inffus.2021.02.022

Zhang, X., Zhang, F., & Xu, C. (2023). VQACL: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19102–19112).

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint* arXiv:2010.04159.