Visual Question Answering Using BERT And VIT Family Models

Using the power of Textual and Vision transformers for visual question and answering

Subbhashit Mukherjee
MT2023065
IIIT Bangalore
Bengaluru, India
Subbhashit.Mukherjee@iiitb.ac.in

Chittaranjan Chandwani
MT2023193
IIIT Bangalore
Bengaluru, India
Chittaranjan.Chandwani@iiitb.ac.in

Nikhil Gupta
MT2023187
IIIT Bangalore
Bengaluru, India
Nikhil.Gupta@iiitb.ac.in

***Abstract*—This paper investigates the combination of bert Bidirectional Encoder Representations from Transformers (BERT) family models with vision transformer (ViT) family models to enhance Visual Question Answering (VQA) systems. VQA requires models to understand and reason about visual content in tandem with natural language questions. Our approach uses a cross-modal attention mechanism to effectively merge visual and textual data, allowing the model to focus on relevant image regions based on the question context. This integration aims to improve both accuracy and interpretability over traditional methods that handle visual and textual components separately. This work demonstrates the potential of combining BERT and (ViT) models for robust VQA performance offering a promising direction for future multi-modal ai system.**

Keywords— BERT, Vision Transformer (ViT), Cross-modal attention, Multi-modal AI, Natural language processing.

I. Introduction

Visual Question Answering (VQA) is a task with lot of challenges that lies at the intersection of computer vision and natural language processing (NLP). It requires models to comprehend both visual content and textual questions to generate accurate answers. The ability to effectively reason about multi-modal data is crucial for developing intelligent systems capable of understanding and responding to human queries in diverse scenarios. Traditional approaches to VQA have often relied on handcrafted features or separate processing pipelines for visual and textual data, limiting their ability to capture complex interactions between these modalities.Recent advancements in deep learning, particularly the emergence of transformer-based models, offer promising avenues for addressing this challenge[4].

Models such as BERT (Bidirectional Encoder Representations from Transformers) and its variants have revolutionized NLP tasks by pre-training on large text corpora and fine-tuning on downstream tasks, achieving state-of-the-art performance in various language understanding tasks. Similarly, Vision Transformer (ViT), Data-efficient Image Transformers (DeiT), and Bidirectional Encoder Representations from Image Transformers (BEiT) have demonstrated remarkable success in image recognition tasks by applying transformer architectures directly to image patches, eliminating the need for handcrafted features or convolutional layers [2].

In this paper, we explore the integration of multiple transformer models for Visual Question Answering. We utilize a combination of ViT, DeiT, and BEiT for image processing, and BERT and RoBERTa for language understanding. By leveraging the strengths of these diverse models and employing a cross-modal attention mechanism, we aim to enhance the performance of VQA systems.

Our approach enables dynamic interactions between visual and textual representations, allowing the model to focus on relevant image regions while processing the accompanying question. Additionally, we apply the Low-Rank Adaptation (LoRA) algorithm to these models to further improve their adaptability and performance.This comprehensive evaluation demonstrates the potential of combining various image and language transformers with LoRA for robust VQA performance, offering a promising direction for future multi-modal AI systems[5].

II. Dataset Creation

For this research, we employed the Visual Question Answering (VQA) v2 balanced image dataset, a comprehensive collection that pairs images with questions designed to gauge the model's understanding and reasoning capabilities. The dataset was meticulously curated to include train and validation questions stored in JSON files, each linked to corresponding image IDs. Additionally, we utilized an annotations JSON file, which provided multiple answers for each question ID and associated image ID, ensuring a robust ground truth for training and evaluation

Our initial step involved renaming the image files to a more suitable format for processing. For example, an image originally named COCO_train2014_000000000025.jpg was renamed to image25. This renaming convention facilitated easier handling and integration of image data with the corresponding questions and annotation [1]s. We did apply the process separately for both train images directory and validation images directory. We also extratcted the names of all the images present in the both the directories so that we can check for the valid pairs in our dataset. By valid pairs , we mean that for all the images present in the dataframe, they must also be present in one of the directories.

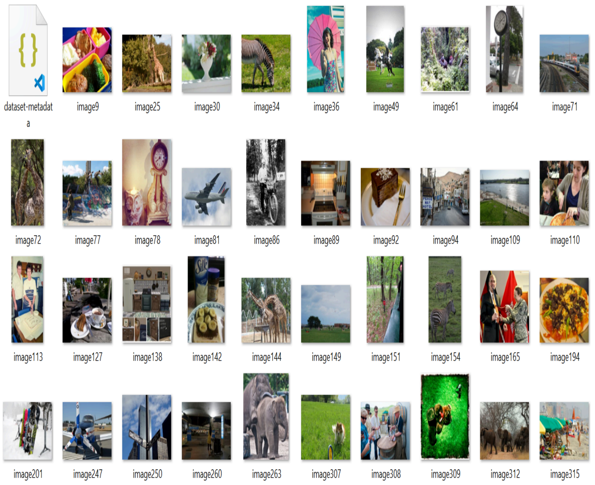
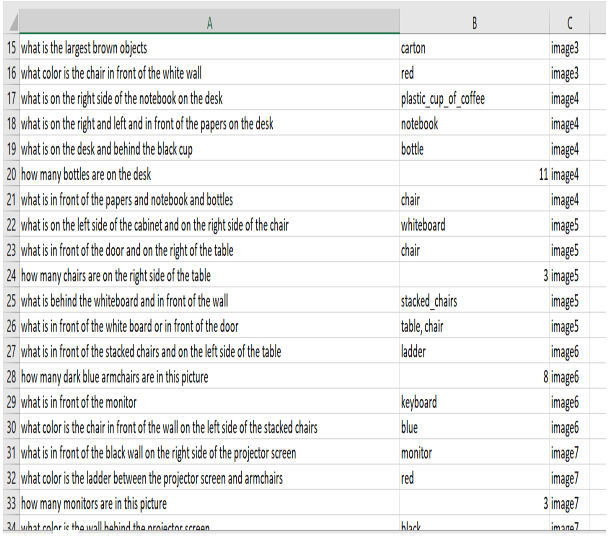


Fig. 1. Image directory after processing

Following this, we created a comprehensive dataset by combining each question with its possible answers for every image, generating separate datasets for training and validation phases. This systematic approach ensured that each question-answer pair was correctly mapped to its respective image, forming the basis for subsequent model training.To ensure data integrity and relevance, we removed any rows in which the image might not be present in the designated image directory. This step was crucial in maintaining a clean and consistent dataset, free from discrepancies that could affect model performance. From the obtained dataset, we randomly selected 10000 rows from it and stored it in a file called data.csv.



	A	B	C
15	what is the largest brown objects	carton	image3
16	what color is the chair in front of the white wall	red	image3
17	what is on the right side of the notebook on the desk	plastic_cup_of_coffee	image4
18	what is on the right and left and in front of the papers on the desk	notebook	image4
19	what is on the desk and behind the black cup	bottle	image4
20	how many bottles are on the desk	11	image4
21	what is in front of the papers and notebook and bottles	chair	image4
22	what is on the left side of the cabinet and on the right side of the chair	whiteboard	image5
23	what is in front of the door and on the right of the table	chair	image5
24	how many chairs are on the right side of the table	3	image5
25	what is behind the whiteboard and in front of the wall	stacked_chairs	image5
26	what is in front of the white board or in front of the door	table, chair	image5
27	what is in front of the stacked chairs and on the left side of the table	ladder	image6
28	how many dark blue armchairs are in this picture	8	image6
29	what is in front of the monitor	keyboard	image6
30	what color is the chair in front of the wall on the left side of the stacked chairs	blue	image6
31	what is in front of the black wall on the right side of the projector screen	monitor	image7
32	what color is the ladder between the projector screen and armchairs	red	image7
33	how many monitors are in this picture	3	image7
34	what color is the wall behind the projector screen	black	image7

Fig. 2. File containing question answer and image pairs

The refined datasets were then split into training and validation sets, which were saved as data_train.csv and data_eval.csv, respectively. This clear demarcation facilitated structured and efficient training and evaluation processes.Additionally, we compiled a text file containing all possible answers, sorted them in a dataframe, and stored them in a file named answer_space.txt. This file served as a reference for the model, providing a comprehensive vocabulary of potential answers, thereby enhancing its ability to accurately predict responses. Finally, all processed files, including the images directory, were uploaded to Kaggle. This platform provided an optimal environment for further modeling, leveraging its computational resources and collaborative features to advance our research effectively.

III. Transformers

This section deals with the extraction of features using transformers. For the purpose of Visual Question-Answering, we need to extract features from both Images and the question-answer space (language). Hence we have implemented a variety of combinations of various language and image transformers.

A. Language Transformers

The introduction of new age language transformers has broadened the scope of natural language processing (NLP), thereby helping textual models to capture the complex dependencies and contexts across long text sequences. Such advancements have significantly improved the performance of various applications of NLP, such as machine translation, text summarization, and giving answers to asked questions. By processing entire text sequences at once, transformers can model intricate linguistic relationships more effectively than traditional methods.A groundbreaking model in this evolution is Bidirectional Encoder Representations from Transformers (BERT). BERT utilizes a unique bidirectional approach, examining the text by considering both the words before and after a given token to fully grasp its context. This dual-directional processing enables BERT to create highly detailed and sophisticated text representations[6]. In our research, we leverage BERT to extract nuanced features from questions and answers, substantially enhancing the depth and accuracy of our visual question answering (VQA) models. BERT's extensive pre-training on large text corpora, followed by fine-tuning for specific tasks, allows it to capture a broad range of language features, making it highly versatile and effective.

Moreover, we have integrated RoBERTa (Robustly optimized BERT approach) into our models. RoBERTa builds on the BERT architecture with key refinements in its training methodology. Unlike BERT, which uses both the Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks, RoBERTa focuses exclusively on MLM. This focused approach allows RoBERTa to train on larger datasets and longer text sequences, leading to improved performance. RoBERTa's dynamic masking and capability to handle larger batch sizes enhance its robustness and efficiency. These attributes make RoBERTa particularly effective in processing extensive data and generating strong text representations, which greatly benefit our VQA models. We have also employed Low-Rank Adaptation (LoRA) to further optimize our transformer models. LoRA enhances efficiency by reducing the number of parameters required during training, facilitating faster and more cost-effective fine-tuning. Incorporating LoRA with both BERT and RoBERTa enables us to maintain high performance while optimizing resource use. This approach is particularly advantageous for our VQA models, as it allows us to manage large-scale data more efficiently and quickly. The combination of LoRA's parameter efficiency with the contextual strengths of BERT and RoBERTa creates a powerful framework for feature extraction. Integrating BERT, RoBERTa, and LoRA in our VQA models capitalizes on their individual strengths in capturing contextual information and optimizing resources. BERT’s bidirectional context processing, RoBERTa's refined training for larger datasets, and LoRA's efficient parameter usage collectively provide a robust solution for feature extraction. By utilizing these models together, we aim to enhance the accuracy and robustness of our VQA systems. The sophisticated representations produced by these transformers, further improved by LoRA, are essential for accurately interpreting and responding to visual questions, which often require an intricate blend of visual and textual data.

B. Image Transformers

Image transformers leverage the principles of transformers and attention mechanisms, originally designed for language tasks, to address computer vision challenges. Rather than converting images into tokens, image transformers decompose images into patches, which are subsequently transformed into vectors. These vectors are processed by transformers to encode images, enabling various computer vision applications, including Visual Question Answering (VQA).

The Vision Transformer (ViT) was among the pioneering models to adapt the transformer architecture for image recognition tasks. It segments an image into fixed-size patches, each linearly embedded, and applies transformer layers to these embeddings. ViT's capability to capture global context through self-attention mechanisms is essential for understanding the interrelationships within an image, crucial for answering questions about the image. In VQA, this allows the model to interpret complex visual scenes and provide accurate responses based on integrated information from various patches.Data-efficient Image Transformers (DeiT) extend the ViT framework by incorporating several strategies to enhance training efficiency, making transformers more accessible without the need for vast datasets[3]. Techniques such as knowledge distillation, where a smaller transformer (student) learns from a larger, pre-trained model (teacher), are employed. DeiT's efficiency and improved training methods enable it to achieve high performance even with smaller datasets, making it suitable for VQA tasks where annotated data may be limited. This makes DeiT particularly valuable in practical VQA applications where extensive labeled data is not always available.

Bidirectional Encoder Representation from Image Transformers (BeIT) adapts the masked language modeling strategy from BERT to images. It involves masking parts of the image and training the model to predict these masked segments, allowing the model to learn rich bidirectional representations. BeIT's pre-training approach facilitates the learning of robust image representations, which can be fine-tuned for VQA tasks. The model's ability to infer and complete missing information is especially advantageous for answering detailed questions about images, as it can deduce context and provide more accurate answers based on incomplete visual data.

In VQA, these image transformers bridge the gap between visual inputs and textual questions, enabling more accurate and context-aware responses. Their capability to manage global context and intricate details in images renders them powerful tools for VQA applications. By leveraging the strengths of models such as ViT, DeiT, and BeIT, VQA systems can achieve higher accuracy and offer more insightful answers, enhancing their applicability in real-world scenarios.

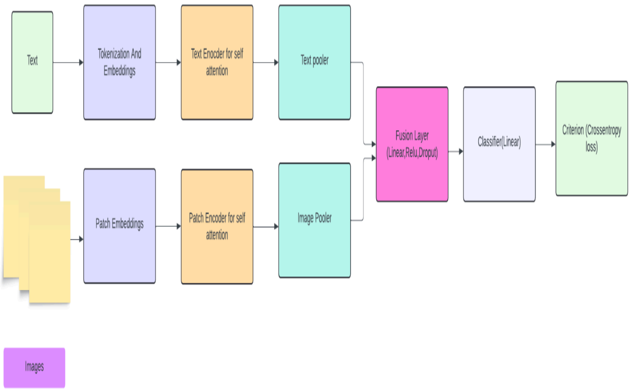

Text
Tokenization And Embeddings
Text Enocder for self attention
Text pooler
Fusion Layer (Linear,Relu,Droput)
Classifier(Linear)
Criterion (Crossentropy loss)
Patch Embeddings
Patch Encoder for self attention
Image Pooler
Images

Fig. 3. Architecture Used

Here is the architecture that we have used for making predictions. It combines the power of both textual-based transformers and image transformers as mentioned above.

IV. Results and Analysis

In this paper we have utilized the power of BERT family models, such as BERT and RoBERTa, along with Vision Transformer,i.e. (ViT) family models, including ViT, BeIT, and DeiT, to make predictions answering for questions that are asked about a given image. This integrated technique combines the strengths of both language and vision transformers, enhancing the accuracy and contextual relevance of the responses generated in Visual Question Answering (VQA) tasks.

By combining these models, our application benefits from the robust language understanding capabilities of BERT and RoBERTa, which excel in preprocessing and understanding the actual textual data, having complex sentence structures and semantics. This ensures us that the questions that were asked about the images are answered correctly and that the nuances of language are effectively captured. On the other hand, vit family models like ViT, BeIT, and DeiT bring advanced techniques and visual processing capabilities to our application. ViT's ability of understanding of global context through powerful self-attention mechanisms helps in identifying complex relationships within different parts of an image. Bidirectional representation of BeIT'sand the ability to predict masked parts of an image heavily contribute to understanding incomplete visual data. DeiT's efficiency in training and performance with smaller datasets makes it adaptable and effective even in scenarios with limited annotated data.

A. BERT AND OTHER VISION TRANSFORMERS

B.

Here , we have generated all the possible combinations of bert transformer and all other vision transformers mentioned above in the paper and reported the metrics as shown below:

Table 1. Bert Familly models accuracy and f1 score

NAME	ACCURACY	F1_SCORE
BERT + VIT	0.267	0.039
BERT + VIT (LORA RANK = 32)	0.226	0.031
BERT + VIT (LORA RANK = 64)	0.207	0.024
BERT + DeIT	0.266	0.038
BERT + DeIT(LORA RANK = 32)	0.233	0.032
BERT + DeIT(LORA RANK = 64)	0.225	0.028
BERT + BeIT	0.260	0.037
BERT + BeIT(LORA RANK = 32)	0.225	0.027
BERT + BeIT(LORA RANK = 64)	0.216	0.025

Table 2. Bert Family models precision, recall and time taken

NAME	PRECISION	RECALL	TIME TAKEN (minutes)
BERT + VIT	0.042	0.034	65.05
BERT + VIT (LORA RANK = 32)	0.038	0.031	69.30
BERT + VIT (LORA RANK = 64)	0.039	0.024	72.701
BERT + DeIT	0.039	0.046	69.23
BERT + DeIT(LORA RANK = 32)	0.035	0.039	65.02
BERT + DeIT(LORA RANK = 64)	0.032	0.038	67.43
BERT + BeIT	0.041	0.042	65.45
BERT + BeIT(LORA RANK = 32)	0.036	0.035	68.90
BERT + BeIT(LORA RANK = 64)	0.031	0.029	65.87

From here we can see that BERT and VIT gave us the highest accuracy whereas BERT and VIT (LORA RANK = 64) gave us the lowest accuracy from the bert transformer combinations.

C. ROBERTA AND OTHER VISION TRANSFORMERS

Here , we have generated all the possible combinations of bert transformer and all other vision transformers mentioned above in the paper and reported the metrics as shown below:

Table 3. Roberta Familly models accuracy and f1 score

NAME	ACCURACY	F1_SCORE
RoBERTa + VIT	0.271	0.031
RoBERTa + VIT (LORA RANK = 32)	0.259	0.028
RoBERTa + VIT (LORA RANK = 64)	0.221	0.0347
RoBERTa + DeIT	0.261	0.033
RoBERTa + DeIT(LORA RANK = 32)	0.235	0.029
RoBERTa + DeIT(LORA RANK = 64)	0.218	0.031
RoBERTa + BeIT	0.260	0.033
RoBERTa BeIT(LORA RANK = 32)	0.241	0.032
RoBERTa + BeIT(LORA RANK = 64)	0.239	0.031

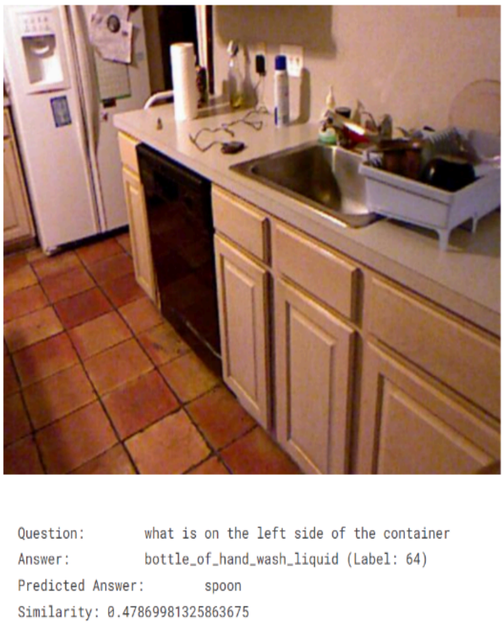
Table 4. Roberta Family models precision, recall and time taken

NAME	PRECISION	RECALL	TIME TAKEN (minutes)
RoBERTa VIT	0.030	0.035	66.54

RoBERTa + VIT (LORA RANK = 32)	0.034	0.029	72.73
RoBERTa + VIT (LORA RANK = 64)	0.037	0.032	72.53
RoBERTa + DeIT	0.031	0.044	77.05
RoBERTa + DeIT(LORA RANK = 32)	0.029	0.035	77.70
RoBERTa + DeIT(LORA RANK = 64)	0.034	0.036	77.18
RoBERTa + BeIT	0.038	0.038	65.99
RoBERTa + BeIT(LORA RANK = 32)	0.035	0.039	65.91
RoBERTa + BeIT(LORA RANK = 64)	0.033	0.032	69.23

From here we can see that Roberta and VIT gave us the highest accuracy whereas BERT and DeIT(LORA RANK = 64) gave us the lowest accuracy from the roberta transformer combinations.

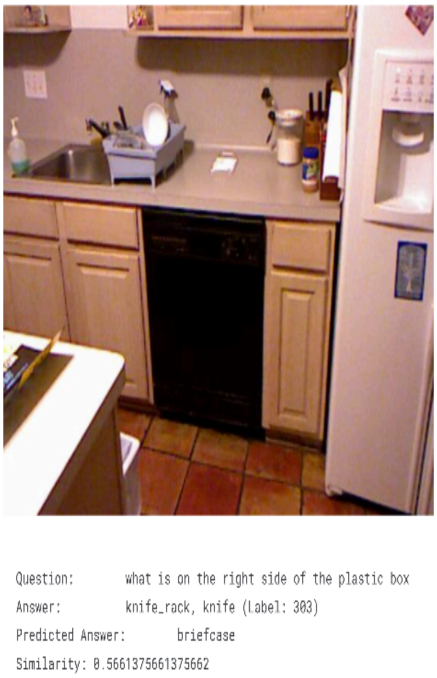
Now we have shown some outputs generated by our models for some random images from our test dataset which we generated by ourselves. Here is the output screenshot of BERT AND DeIT transformer predicting the answer with a similarity score of 0.47 for the question asked for this image.



```
Question:        what is on the left side of the container
Answer:          bottle_of_hand_wash_liquid (Label: 64)
Predicted Answer:        spoon
Similarity: 0.47869981325863675
```

Fig. 2. BERT + DEIT WITHOUT LORA

We also utilized lora on the same configuration and shown some outputs generated by our models for some random images from our test dataset which we generated by ourselves. Here is the output screenshot of BERT AND DeIT transformer along with lora configuration for predicting the answer .We got a similarity score of 0.56 for the question asked for this image.



```
Question:        what is on the right side of the plastic box
Answer:          knife_rack, knife (Label: 303)
Predicted Answer:        briefcase
Similarity: 0.5661375661375662
```

Fig. 2. BERT + DEIT WITH LORA

In the end , we were able to provide pretty good answers for the specified question for an image.

Conclusion

In the Visual Question Answering (VQA) project, the combination of BERT and ViT emerged as the most balanced and highest-performing configuration. This model demonstrated the best overall performance with an accuracy of 0.267 and an F1 score of 0.039. Additionally, it maintained a reasonable balance between precision (0.042) and recall (0.034) while requiring a moderate computational time of 65.05 minutes.

On the other hand, the RoBERTa and ViT combination achieved the highest accuracy at 0.271, but this came at the expense of a lower F1 score (0.031), indicating a less reliable balance between precision and recall. This suggests that while RoBERTa and ViT can be highly accurate, they might not perform as consistently across different aspects of the VQA task.

The introduction of LoRA (Low-Rank Adaptation) ranks, intended to enhance model performance by reducing the parameter space, generally resulted in reduced performance metrics and increased computation time.As we have included only 10000 pairs of questions and answers, this might be an issue. Although these configurations did not improve the performance in this particular project, they might still be valuable depending on specific resource constraints and acceptable performance trade-offs.

The primary models used in this project—BERT, RoBERTa, ViT, DeiT, and BeIT—are already highly optimized and complex. The limited parameter space introduced by LoRA might have constrained the models' ability to learn and generalize effectively, especially if the dataset was not sufficiently large or diverse. This highlights the importance of considering the specific characteristics and requirements of the dataset and task when choosing and optimizing model configurations.

Overall, while BERT + ViT provided the best balance of performance and computational efficiency, the choice of model configuration should be tailored to the specific needs and constraints of the project. The insights gained from this analysis can guide future optimizations and adjustments to achieve the desired balance of accuracy, precision, recall, and computational requirements in VQA applications.

References

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[3] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR.

[4] 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).

[5] Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3-20.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.arXiv preprint arXiv:1907.11692.