Visual Question Answering

G Sai Hemanth Kumar
MT2023010
IIIT Bangalore
Bengaluru, India
Sai.Hemanth@iiitb.ac.in

Billa Abhignan
MT2023044
IIIT Bangalore
Bengaluru, India
Billa.Abhignan@iiitb.ac.in

Nabaneet Dutta Kanungoe
MT2023194
IIIT Bangalore
Bengaluru, India
nabaneet.kanungoe@iiitb.ac.in

***Abstract*— This paper presents a novel approach to enhancing Visual Question Answering (VQA) systems by integrating advanced pre-trained models and methodologies. We focus on fine-tuning state-of-the-art pre-trained models, specifically the Facebook DPR question encoder and the Vision Transformer (ViT) for image feature extraction. Our methodology involves augmenting the VQA dataset with enriched embeddings from the Facebook DPR question encoder and utilizing the ViT to extract image features, thereby improving data quality. We propose a hybrid architecture combining dense layers for image feature extraction with custom question encoding mechanisms. Through rigorous experimentation, we evaluate our approach, providing insights into personalized VQA systems. Our findings contribute to advancing VQA technology by leveraging state-of-the-art pre-trained models, ultimately enhancing multimedia interaction experiences**

Index terms — VQA, Pre-trained models, Facebook DPR question encoder, ViT, Image feature extraction, Hybrid architecture, Multimedia interaction

I. Introduction

In today's dynamic market landscape, the integration of image understanding and natural language processing has become increasingly pivotal, particularly in applications such as Visual Question Answering (VQA). This burgeoning field intersects the realms of computer vision and natural language understanding, aiming to develop systems capable of comprehending and responding to questions posed about visual content. With the proliferation of multimedia platforms and the growing demand for interactive technologies, VQA has emerged as a cornerstone in enhancing user engagement and facilitating seamless human-computer interaction.

However, despite the potential benefits, the development of robust VQA systems faces several challenges. One major hurdle is the inherent complexity of bridging the semantic gap between images and textual queries. Images contain rich, unstructured information, making it challenging to extract relevant features and comprehend their semantic context accurately. Likewise, natural language queries exhibit nuanced semantics, requiring models to possess a deep understanding of language structure and meaning. Addressing these challenges necessitates the development of sophisticated algorithms capable of effectively integrating visual and textual modalities.

In recent years, the adoption of pre-trained models has revolutionized the landscape of machine learning and natural language processing. These models, pre-trained on vast amounts of data, possess rich representations of linguistic and visual knowledge, offering a promising foundation for various downstream tasks. Leveraging pre-trained models such as the Facebook DPR question encoder and DINO for image feature extraction enables practitioners to bootstrap their models with valuable knowledge and significantly reduce the need for large-scale annotated datasets.Additionally, innovative combinations like DIET+BERT have been explored to further enhance the performance of models by integrating robust intent recognition with deep contextual understanding.

The fine-tuning of pre-trained models represents a powerful strategy for adapting these models to specific tasks or domains. By fine-tuning pre-trained models on task-specific datasets, practitioners can enhance model performance and adapt them to the intricacies of the target task. However, fine-tuning presents its own set of challenges, including the risk of overfitting, the need for domain-specific data, and the delicate balance between retaining useful features and adapting to new information.

In this study, we explore the efficacy of fine-tuning pre-trained models, both with and without the use of techniques such as LORA, in the context of Visual Question Answering. We investigate the impact of fine-tuning on model performance, addressing challenges related to data scarcity, model adaptation, and interpretability. Additionally, we examine how leveraging pre-trained models enhances the depth and quality of input data, paving the way for more sophisticated and effective VQA systems. Through rigorous experimentation and comparative analysis, we aim to elucidate the potential of fine-tuning pre-trained models for advancing the field of VQA and enriching the multimedia interaction experience.

II. Architecture

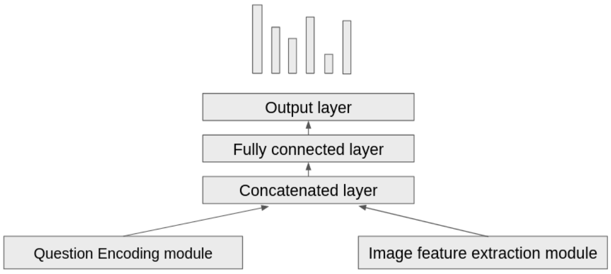


Fig. 1. Model Architecture without LoRA

The architecture begins with the Question Encoding Module, which takes textual questions as input. This module encodes questions into dense embeddings, capturing the semantic information of the questions. These embeddings facilitate better understanding and integration with image features. The Image Feature Extraction Module processes images to complement the textual input. This module extracts high-dimensional feature representations from images, capturing visual semantics.

The Concatenation Layer merges the encoded question embeddings with the extracted image features. By combining textual and visual information, this layer enhances the model's understanding of the question-image pairs.

Following the concatenation layer, the architecture incorporates dense layers for feature fusion and abstraction. These dense layers facilitate the integration of multi-modal information and capture complex relationships between textual and visual inputs. The Output Layer receives the fused representations and performs classification to predict the answer to the input question. This layer utilizes a softmax activation function to produce probability distributions over the possible answers. The predicted answer is selected based on the highest probability.

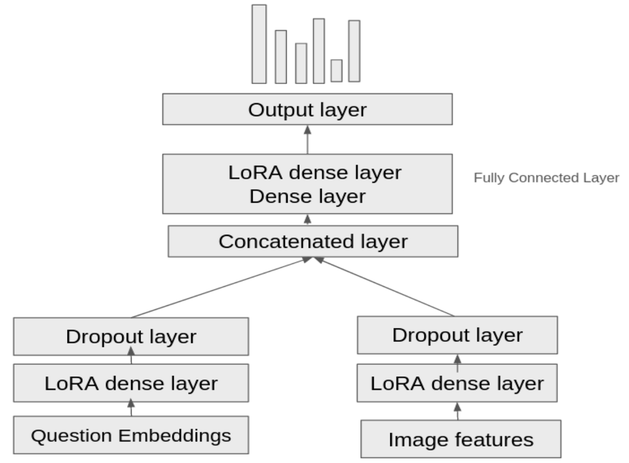


Fig. 2. Model architecture with LoRA

The architecture begins with the Question Encoding Module, which encodes textual questions into dense embeddings, capturing their semantic information. The Image feature extraction module extracts high-dimensional feature representations from images, capturing visual semantics. Before concatenation, dropout layers are applied separately to both the encoded question embeddings and the extracted image features. Dropout layers help in regularizing the model, preventing overfitting by randomly setting a fraction of input units to zero during training.

The Concatenation Layer merges the dropout-regularized question embeddings with the extracted image features. Following the concatenation layer, the architecture incorporates dense layers for feature fusion and abstraction. The Output Layer receives the fused representations and performs classification to predict the answer to the input question. This layer utilizes a softmax activation function to produce probability distributions over the possible answers. The predicted answer is selected based on the highest probability.

The integration of LoRA layers in the dense layers significantly improves the model's efficiency and adaptability. LoRA allows for efficient fine-tuning with fewer trainable parameters, enhancing the model's ability to fuse textual and visual features. This innovative approach enables the model to generate accurate answers to questions posed about images. Through fine-tuning and experimentation, the architecture with LoRA aims to optimize performance, improve parameter efficiency, and advance the field of VQA.

III. Experimentation

We applied two distinct methodologies to tackle the VQA task: firstly, the Diet + BERT framework, integrating robust intent recognition with deep contextual understanding; and secondly, the DINO + DPR framework, leveraging self-supervised learning with dense passage retrieval for enhanced understanding of both image and text modalities.

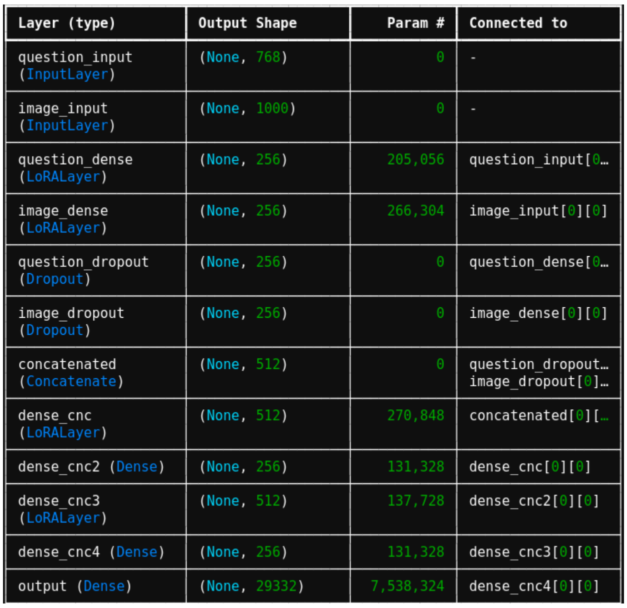
A. DIET+BERT

We utilize the Diet (Data-efficient Image Transformers) model for image feature extraction. Diet provides efficient and effective image representations, crucial for understanding visual content. These features are extracted for each image in the dataset, resulting in a high-dimensional feature vector. For encoding textual questions, we employ the BERT (Bidirectional Encoder Representations from Transformers) model. BERT is adept at capturing the intricate contextual nuances of language, making it ideal for understanding the semantics of questions. The questions are tokenized and encoded using BERT, yielding dense embeddings that encapsulate their meaning.

The Diet + BERT model is trained using a combination of stochastic gradient descent (SGD) with momentum optimization and categorical cross-entropy loss. Training is conducted over multiple epochs, with careful monitoring of validation accuracy to prevent overfitting. The best performing model is selected based on its ability to accurately answer questions from the validation set.

We introduce LoRA layers to facilitate enhanced interaction between the question and image features. LoRA layers enable the model to focus on relevant parts of the input space, promoting better fusion of visual and textual information. This attention mechanism aids in capturing fine-grained correlations between images and questions, leading to improved VQA performance.

Upon evaluation, the Diet + BERT model demonstrates promising results, showcasing its efficacy in handling the VQA task. By leveraging state-of-the-art pre-trained models and innovative architectural enhancements like LoRA, our approach achieves notable improvements in accuracy and generalization capability.



Layer (type)	Output Shape	Param #	Connected to
question_input (InputLayer)	(None, 768)	0	-
image_input (InputLayer)	(None, 1000)	0	-
question_dense (LoRALayer)	(None, 256)	205,056	question_input[0…
image_dense (LoRALayer)	(None, 256)	266,304	image_input[0][0]
question_dropout (Dropout)	(None, 256)	0	question_dense[0…
image_dropout (Dropout)	(None, 256)	0	image_dense[0][0]
concatenated (Concatenate)	(None, 512)	0	question_dropout… image_dropout[0]…
dense_cnc (LoRALayer)	(None, 512)	270,848	concatenated[0][…
dense_cnc2 (Dense)	(None, 256)	131,328	dense_cnc[0][0]
dense_cnc3 (LoRALayer)	(None, 512)	137,728	dense_cnc2[0][0]
dense_cnc4 (Dense)	(None, 256)	131,328	dense_cnc3[0][0]
output (Dense)	(None, 29332)	7,538,324	dense_cnc4[0][0]

Fig. 3. Summary with LoRA

B. *DINO+DPR*

For the experimentation using the DINO + DPRmodel, we delved into the integration of advanced pre-trained models and methodologies to enhance the efficacy of VQA systems. Questions were tokenized and padded to ensure uniformity in input dimensions for model compatibility.Answers were encoded into one-hot vectors, providing a categorical representation of the possible answer choices.

Image features were extracted using the DINO (VIT) model, resulting in high-dimensional feature representations capturing visual semantics efficiently. The dataset was partitioned into training, validation, and testing sets, with a respective allocation of 80%, 10%, and 10% of the total data. To enhance the robustness of the model, data augmentation techniques such as rotation, flipping, and scaling were employed to diversify the training samples. Additionally, a portion (25%) of the images was randomly selected and augmented to augment the dataset's diversity further.

The DINO + DPR architecture comprised a fusion of the DINO-based image features and enriched embeddings generated by the Facebook DPR question encoder. Image features and question embeddings were concatenated and fed into a hybrid architecture consisting of dense layers for feature fusion and abstraction.The model was trained using a custom-designed optimizer, aiming to minimize categorical cross-entropy loss and maximize prediction accuracy. Training was conducted over multiple epochs, with hyperparameters tuned based on performance metrics evaluated on the validation set.

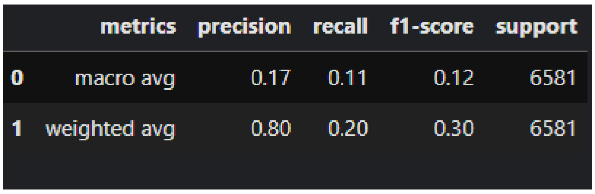
To further enhance the model's performance and efficiency, we integrated LoRA (Low-Rank Adaptation) layers into the dense layers for feature fusion and abstraction. This approach not only enhances the fusion of textual and visual features but also contributes to reducing overfitting and increasing the overall robustness of the model. Through careful tuning and experimentation, the inclusion of LoRA layers has demonstrated improved efficiency and effectiveness in generating accurate answers to questions posed about images, advancing the field of VQA.

IV. Results and Analysis

A. *DIET+BERT*

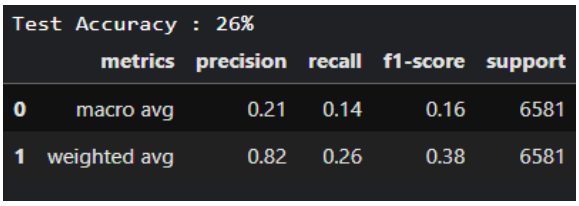
In this section, we analyze the results obtained from our experimentation using the DIET + BERT model for VQA. We discuss the precision, recall, and F1-score metrics to provide a comprehensive evaluation of the model's performance. We focus on the comparison of validation accuracy across different configurations: with LoRA (rank=1), with LoRA (rank=8), and without LoRA.

We evaluated the model using precision, recall, and F1-score metrics, summarized in the following table:



	metrics	precision	recall	f1-score	support
0	macro avg	0.17	0.11	0.12	6581
1	weighted avg	0.80	0.20	0.30	6581

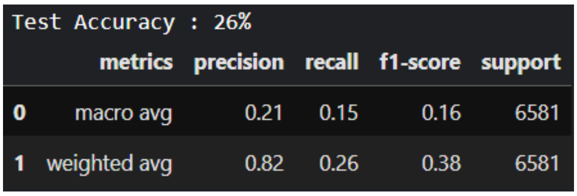
Fig. 4. Precision,Recall and F1-Score without LoRA



Test Accuracy : 26%

	metrics	precision	recall	f1-score	support
0	macro avg	0.21	0.14	0.16	6581
1	weighted avg	0.82	0.26	0.38	6581

Fig. 5. Precision,Recall and F1-Score with LoRA (rank=1)



Test Accuracy : 26%

	metrics	precision	recall	f1-score	support
0	macro avg	0.21	0.15	0.16	6581
1	weighted avg	0.82	0.26	0.38	6581

Fig. 6. Precision,Recall and F1-Score with LoRA (rank=8)

“Macro Average” indicates an average performance across all classes, treating each class equally. “Weighted average” accounts for the number of instances in each class, providing a more holistic view of the model's performance.

Validation accuracy over epochs:

The plot illustrates the validation accuracy across epochs for three configurations: LoRA (rank=1), LoRA (rank=8), and without LoRA. The model without LoRA demonstrates a steady increase in validation accuracy from 23.85% at epoch 6 to 26.25% at epoch 15. LoRA (rank=1) shows superior performance compared to the baseline. It starts at 24.54% validation accuracy at epoch 6 and reaches 27.49% by epoch 15. LoRA (rank=8) also improves over the baseline but at a slower rate compared to LoRA (rank=1). It starts at 25.37% at epoch 6 and reaches 28.07% by epoch 15.

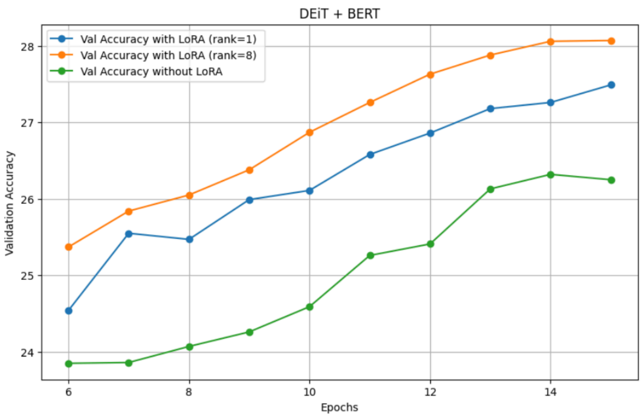


Fig. 7. Validation accuracies over epochs

Validation accuracy over time:

We plotted the validation accuracy against the training time. Each epoch corresponds to 10 minutes of training. The model without LoRA shows a gradual increase in validation accuracy from 23.85% at 60 minutes (epoch 6) to 26.25% at 150 minutes (epoch 15). LoRA (rank=1) starts at 24.54% validation accuracy at 60 minutes and reaches 27.49% at 150 minutes. LoRA (rank=8) begins at 25.37% at 60 minutes and increases to 28.07% at 150 minutes.

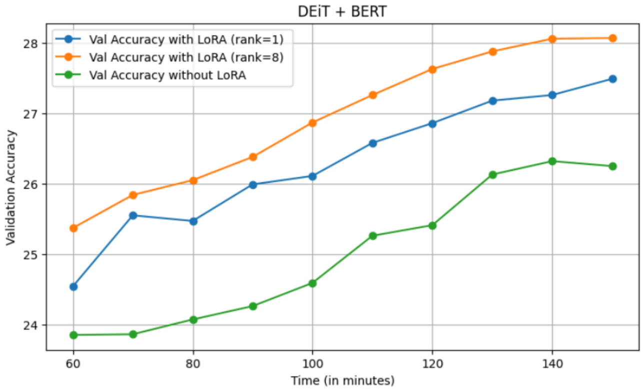


Fig. 8. Validation accuracies over time

These results highlight the effectiveness of LoRA in improving validation accuracy, with LoRA (rank=1) showing the most significant improvement over the baseline model. LoRA (rank=8) also enhances performance but at a slower rate compared to LoRA (rank=1). The integration of LoRA layers into the DIeT + BERT model demonstrates their potential in enhancing the fusion of visual and textual information for better VQA performance.

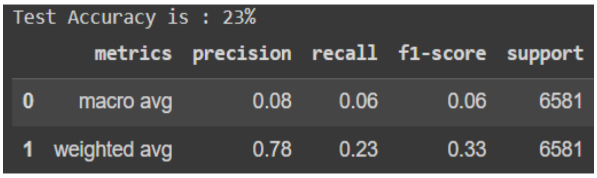
For complex tasks requiring detailed understanding and generation, higher ranks are usually beneficial as they provide the necessary capacity for the model to adapt appropriately. Since BERT is a generalized model it doesn’t capture question embeddings as well as it should. Therefore a higher rank (rank=8) facilitates to capture much greater complexity than rank=1 that shows a better accuracy for rank=8.

“Rank” parameter directly influences the model's capacity and efficiency, providing a more intuitive and impactful way to control the adaptation process. Adjusting the rank offers clear benefits in terms of balancing model expressiveness and computational efficiency, making it a preferred choice in practice.

B. DINO+ DPR

In this section, we analyze the results obtained from our experimentation using the DINO + DPR model for VQA. We discuss the precision, recall, and F1-score metrics to provide a comprehensive evaluation of the model's performance. We focus on the comparison of validation accuracy across different configurations: with LoRA (alpha=1), with LoRA (alpha=16), and without LoRA.

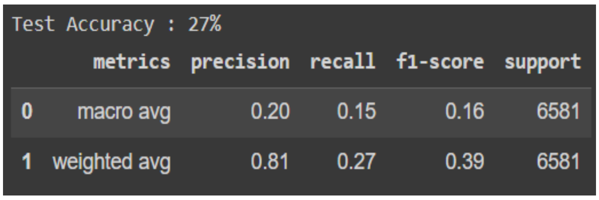
We evaluated the model using precision, recall, and F1-score metrics, summarized in the following table:



Test Accuracy is : 23%

	metrics	precision	recall	f1-score	support
0	macro avg	0.08	0.06	0.06	6581
1	weighted avg	0.78	0.23	0.33	6581

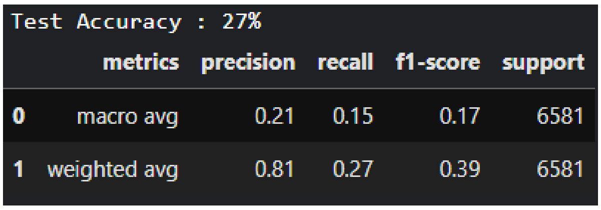
Fig. 9. Precision,Recall and F1-Score without LoRA



Test Accuracy : 27%

	metrics	precision	recall	f1-score	support
0	macro avg	0.20	0.15	0.16	6581
1	weighted avg	0.81	0.27	0.39	6581

Fig. 10. Precision,Recall and F1-Score with LoRA (alpha=1)



Test Accuracy : 27%

	metrics	precision	recall	f1-score	support
0	macro avg	0.21	0.15	0.17	6581
1	weighted avg	0.81	0.27	0.39	6581

Fig.11. Precision,Recall and F1-Score with LoRA (alpha=16)

Validation accuracy over epochs:

The plot illustrates the validation accuracy across epochs for the three configurations: LoRA (alpha=1), LoRA (alpha=16), and without LoRA. The model without LoRA demonstrates a steady increase in validation accuracy from 23.98% at epoch 6 to 27.11% at epoch 15. LoRA (alpha=1) shows superior performance compared to the baseline. It starts at 26.02% validation accuracy at epoch 6 and reaches 29.07% by epoch 15. LoRA (alpha=16) also improves over the baseline but at a slower rate compared to LoRA (alpha=1). It starts at 25.38% at epoch 6 and reaches 28.08% by epoch 15.

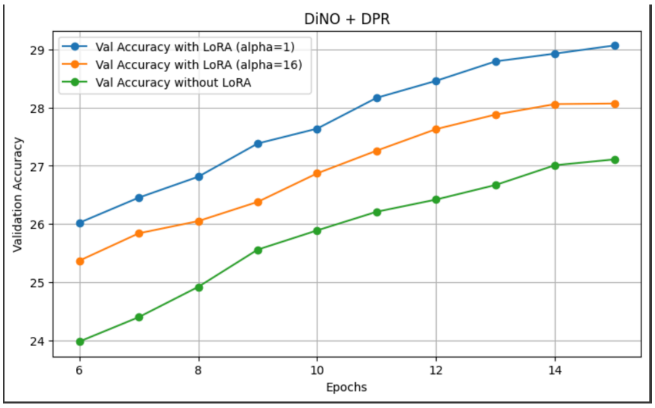


Fig. 12. Validation accuracies over epochs

Overall, the plot clearly demonstrates that integrating LoRA, especially with alpha=1, provides a significant advantage in terms of validation accuracy compared to the model without LoRA.

Validation accuracy over time:

The plot presents the validation accuracy against time (in minutes), offering insights into the efficiency of each configuration. The baseline model without LoRA achieves a validation accuracy of 27.11% after approximately 150 minutes. LoRA (alpha=1) achieves the highest validation accuracy of 29.07% in the same timeframe, illustrating both higher accuracy and efficiency. LoRA (alpha=16) achieves a validation accuracy of 28.08% in about 150 minutes, which is an improvement over the baseline but less efficient compared to LoRA (alpha=1).

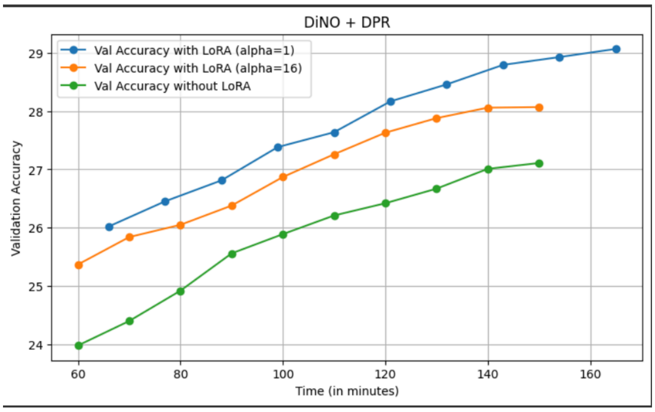


Fig. 13. Validation accuracies over time

This plot indicates that LoRA (alpha=1) not only improves the accuracy but also optimizes the training time, making it a more efficient approach for VQA tasks.

We chose to fine tune the "alpha" parameter for tuning because the "DINO" is well trained on image feature extraction and "DPR" is a specific model trained for question encodings. Alpha controls the magnitude of the updates from the low-rank matrices, it helps in balancing the preservation of pre-trained knowledge with the incorporation of new task-specific adaptations. Accuracy for (alpha=1) is better than accuracy for (alpha=16)

Conclusion

In this study, we explored two advanced methodologies for tackling the VQA task: the Diet + BERT framework and the DINO + DPR framework. The Diet + BERT model, integrating Diet for image feature extraction and BERT for question encoding, demonstrated significant performance improvements with the inclusion of LoRA layers. Specifically, LoRA (rank=1) notably enhanced validation accuracy, showcasing the model's capability to effectively fuse visual and textual information.

Similarly, the DINO + DPR framework, which leverages DINO for visual feature extraction and DPR for dense question retrieval, benefited from LoRA integration. LoRA (alpha=1) provided the best balance between accuracy and training efficiency, outperforming both the baseline and LoRA (alpha=16) configurations. This highlights the importance of fine-tuning specific parameters to optimize model performance.

Overall, the use of LoRA layers in both frameworks significantly boosted VQA accuracy by enhancing the interaction between image and text features. These results demonstrate the potential of combining state-of-the-art pre-trained models with innovative fine-tuning techniques to advance the field of VQA.

Challenge faced: When we try to load entire test dataset, we got an error of "resource exhaust", so we did prediction by taking each data point separately

References

[1] Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing March 20243(1):0311-0320 DOI:10.47813/2782-5280-2024-3-1-0311-0320

[2] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). "Training data-efficient image transformers & distillation through attention." In Proceedings of the 38th International Conference on Machine Learning (ICML).

[3] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). "Emerging properties in self-supervised vision transformers." *IEEE Transactions on Pattern Analysis and Machine Intelligence*

[4] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). "Dense Passage Retrieval for Open-Domain Question Answering."

[5] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). "VQA: Visual Question Answering." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2425-2433). DOI: 10.1109/ICCV.2015.279.

[6] Zhu, Y., Groth, O., Bernstein, M. S., & Fei-Fei, L. (2016). "Visual7W: Grounded Question Answering in Images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4995-5004). DOI: 10.1109/CVPR.2016.540.