LoRA Meets Medical Imagery: Refining BLIP for Enhanced Visual Question Answering

Dinesh Kumar M R, Pillalamarri Akshaya, Saivarsha R, Shrish Surya N T, Premjith B, Sowmya V, and Jyothish Lal G

Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India.
b_premjith@cb.amrita.edu, v_sowmya@cb.amrita.edu, g_jyothishlal@cb.amrita.edu

Abstract. In this work, a LoRA-optimized BLIP model is used to investigate the improvement of Medical Visual Question Answering (VQA). Our work aims to optimize the BLIP architecture through the use of Low-Rank Adaptation (LoRA) to develop a more effective and resource-efficient method for medical image analysis. We rigorously test this technique on a specialized combination of medical VQA datasets and show its efficacy. The outcomes demonstrate notable gains in accuracy, especially for closed-type questions, highlighting the potential of LoRA-enhanced BLIP models to advance AI-driven healthcare solutions and medical diagnostics. This paper lays the groundwork for future research and development in the field by presenting a novel approach that connects cutting-edge AI techniques with essential medical applications.

Keywords: Medical Visual Question Answering, Low-Rank Adaptation (LoRA), Fine-Tuning

1 Introduction

Significant progress has been made in the application of artificial intelligence (AI) to medical diagnostics, especially in the area of visual question answering (VQA). In the specialized field of medical VQA, artificial intelligence (AI) is used to analyze and respond to queries regarding medical images, facilitating diagnosis and treatment planning. Because of its strong multi-modal understanding abilities, the Bootstrapped Language Image Pre-training (BLIP) model has become a powerful tool for VQA in recent times. Still, a challenge is the large amount of computational power needed to train such models.

To tackle this, we present a novel solution in this work: we apply Low-Rank Adaptation (LoRA) to the BLIP model to maximize efficiency while maintaining performance. A recent development in model training called LoRA makes it possible to significantly reduce parameters, which speeds up training and uses fewer resources. Specifically, we target the accuracy and adaptability of the BLIP model in understanding complex medical imagery by fine-tuning it with LoRA

for the Medical VQA task.

This paper outlines our approach, from model adaptation to extensive testing on medical datasets, and shows the improved efficiency and accuracy attained. Our work highlights the potential of effective AI solutions in healthcare by presenting a novel application of LoRA in the field and laying the groundwork for future developments.

2 Related Works

The field of Medical Visual Question Answering (VQA) has seen significant developments, as evidenced by various approaches presented at the VQA-Med 2020 and 2021 events [1]. In 2020, Shengyan showcased an Encoder-Decoder Model [2] leveraging the ImageCLEF VQA-Med 2020 dataset with a combination of VGG16, GRU, and seq2seq, achieving 37.6% accuracy and a score of 0.412. Another notable entry was HARENDRAKV's Sequential VQA with Attention [3], using VGG, BERT, MFB, and GLOVE on the VQA-Med-VQA 2020 dataset, achieving 37.8% accuracy and a score of 0.439. The HCP-MIC's approach [4], focusing on Effective Visual Representation, utilized BioBERT on the ImageCLEF 2020 VQA-Med dataset, resulting in 42.6% accuracy and a score of 0.462.

Bumjun Jung's model [5] based on feature extraction and multi-modal feature fusion, which used VGGNet, GAP, bioBERT, and MFH Pooling with co-attention on the ImageCLEF 2020 VQA-Med dataset, showed impressive results with 46.6% accuracy and a score of 0.502. Additionally, a study titled "LSTM in VQA-Med, is it really needed?" [6] examined the ImageCLEF 2019 dataset using CNN and LSTM, achieving 54% accuracy and a score of 0.55.

In 2021, SSN MLRG's approach [7] for solving abnormality-related queries used the ImageCLEF VQA - MED 2021 dataset with a combination of VGGNet, LSTM, GRU, and BERT, achieving 19.6% accuracy and a score of 0.227. The Inception Team's pre-trained VGG [8] with Data Augmentation method for the ImageCLEF VQA - MED 2020 dataset, using VGG16 and LSTM, achieved 48% accuracy and 0.511 for VQA, and 33.9% accuracy and 0.511 for VQG.

TUA1's approach [9] in 2019, a classification and generation model based on transfer learning using Inception-ResNet-V2, BERT, and LSTM on the VQA-Med2019 dataset, achieved a notable 60.6% accuracy and a score of 0.633. Kdevqa's GLU-based classification [10] in 2020 on the VQA-Med(2020) dataset with VGG16 and BERT achieved 31.4% accuracy and a score of 0.350. Yunnan University's Pretrained BioBERT [11] for the ImageCLEF VQA-Med 2021 achieved 36.2% accuracy and a score of 0.402. TeamS's BBN-Orchestra [12] for Long-tailed Medical VQA in 2021 on the VQA-Med 2021 dataset achieved 34.8% and 39.1% accuracy. Lijie's Attention Model-based [13] Efficient Interaction between Multimodality in 2021 used VGG8 and BioBERT on the ImageCLEF VQA-Med

2020 dataset, achieving 31.6% accuracy and a score of 0.352. Finally, Edward's "LoRA: Low-Rank Adaptation of Large Language Models" [14] describes a novel method for using low-rank matrices to adapt large pre-trained language models, such as GPT-3. This method seeks to improve the efficiency of parameterization and computation during the fine-tuning of these large models, enabling tailored modifications without requiring retraining of the full model. By changing only a small portion of the model's parameters, the low-rank adaptation method preserves the pre-trained weights and uses fewer resources during adaptation.

3 Materials & Methodology

3.1 Dataset

We used a customized dataset designed for Medical Visual Question Answering (VQA) in this investigation. The dataset contains a series of questions and answers that go along with a collection of medical photographs. Three essential components make up each data point in the dataset: an image ID, a question about the medical image, and the right response to the inquiry.

X-rays, MRI scans, CT scans, and ultrasound images are among the several medical imaging modalities from which the images in the dataset are drawn. The dataset is both extensive and demanding due to this variability, which guarantees a broad representation of medical imaging events. For convenience, the pictures have distinctive IDs labeled on them.

The questions are formulated to mimic real-world scenarios where a medical professional or an AI system might need to interpret medical imagery. These questions range from identifying the type of imaging modality used to more complex queries about diagnoses or observations in the images.

Answers are provided in a concise format, either as single words or short phrases, to facilitate straightforward evaluation of the model's predictions. The dataset includes both 'open-ended' questions, where a range of answers might be correct, and 'closed-ended' questions, which have a specific correct answer.

To ensure there is no less quantification in data, we made a combination of MED-2019, VQA-RAD, and SLAKE-English Dataset to ensure there's robustness in the dataset. The combination split-up can be seen in table 1.

Table 1. Combination split-up

Dataset Name	Images	QA Pairs
MED-2019	4200	15,292
VQA-RAD	315	2022
SLAKE-English	642	7033
Combined	5157	24,347

3.2 Data preprocessing

The data preparation step was essential for our optimized BLIP model to analyze the medical picture and text data efficiently. This step entailed several crucial procedures to ensure the data was in the right format and quality for the model's training and testing.

Image Processing: To provide consistent model training and inference, a series of changes were applied to every image in the medical dataset to standardize its size and format.

- **Resizing:** Every image underwent a scaling process to get a consistent 384 × 384 pixel size. This scaling was required to meet the BLIP model's input size specifications.
- **Conversion to Tensors:** The images were resized and then transformed into tensors, which are the common format used by deep learning frameworks for image data.
- **Normalization:** Additionally, we used specified mean and standard deviation values to normalize the image tensors. The pixel values are scaled to a range that is better suited for model training during this critical normalization step, which frequently results in faster convergence and greater generalization.

Text Preprocessing: Several preprocessing stages were applied to the questions and answers related to the photos. To prepare the text data for model training and analysis, several steps are essential. The following protocols were put into place:

- **Tokenization:** Tokenization was used for the questions and responses about the photographs. Tokenization is the process of transforming the unprocessed text into a list of tokens (words or subwords) that the model can understand.
- **Tokenizer Utilization:** We made use of the tokenizer from the BLIP model's Processor which turned out to be BERT itself. Tokenizers like this one are made expressly to handle the kinds of language data that the BLIP model requires.
- **Padding and Truncation:** To guarantee consistent length throughout all sequences, the tokenized text underwent padding and truncation. For neural network batch processing, this consistency is essential.

3.3 Data Splitting and Loader

There were training and testing sets inside the dataset. This separation makes it possible to train the model efficiently on a single subset of data (the training set) and to objectively assess the model's performance on data that hasn't been seen yet (the testing set). The split was carefully selected to prevent both overfitting and underfitting while guaranteeing an adequate amount of data for each phase as can be seen in table 2.

Table 2. Datasets split-up

Dataset Type	Number of Images	Number of QA Pairs
Train	4619	22,561
Test	1453	1786

To feed the preprocessed data into the model during training and testing, a custom data loader was created. The data loader supported actions including batching, shuffling, and parallel processing, ensuring efficient and effective data handling.

One of the most important steps in our research was the careful preparation of the text and image data. It was essential to the success of our Medical VQA system because it allowed the refined BLIP model to efficiently learn from and analyze the intricate medical images and related queries.

3.4 Proposed Framework for VQA

Our study offers a novel method for Medical Visual Question Answering (VQA) by employing Low-Rank Adaptation (LoRA), a method that enhances large-scale models for more effective training and performance, to fine-tune the BLIP (Bootstrapped Language Image Pre-training) model.

1. **Model Architecture:** Our methodology is based on the BLIP model, which is well-known for its ability to comprehend and provide unified vision-language tasks. Because this model architecture can process both textual and visual inputs, it is ideally suited for VQA.

   The provided photos are processed by the **Vision Model**. It converts the visual data into a format that works well for analysis in conjunction with the written data. The queries about medical imagery are encoded by the **Text Encoder**. Because this component lacks a pooling layer, the textual input can be understood in more detail. The task of producing responses falls to the **Text Decoder**. It generates an answer that is appropriate for the circumstances by utilizing the encoded question and image embeddings.

2. **LoRA Adaptation:** To address the computational and resource-intensive issue of training large-scale models, we incorporated LoRA into the BLIP model. By adding low-rank matrices to the BLIP model's transformer layers, LoRA makes it possible for pre-trained weights to be efficiently adjusted. By drastically lowering the number of trainable parameters, this method speeds up training and uses fewer resources.

   A neural network is made up of numerous thick layers that multiply matrices. Usually, these layers' weight matrices are full-rank. The pre-trained language models have a low "intrinsic dimension" when it comes to task adaptation, and they can still learn effectively even when they are randomly projected to a smaller subspace. Motivated by this, we conjecture that during adaptation, the updates to the weights likewise possess a low "intrinsic rank".

   We limit the update of a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, by using a low-rank decomposition $W_0 + \Delta W = W_0 + BA$ to describe it, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d, k)$. While A and B have trainable parameters, W_0 is frozen and does not receive gradient updates during training. It should be noted that the same input is used to multiply both W_0 and $\Delta W = BA$, and the corresponding output vectors are summed coordinate-wise. For $h = W_0 x$, our modified forward pass results in:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

Listing 1.1: BLIP Model Configuration and LoRA Adaptation

```
# BLIP is having over 384M parameters
print_trainable_parameters(model)
trainable params: 384672572 || all params: 384672572 ||
    trainable%: 100.0

# Define LoRA configuration for the BLIP VQA model
config = LoraConfig(
    r=16,  # Rank of LoRA
    lora_alpha=32,  # Scaling factor
    lora_dropout=0.1,  # Dropout for LoRA layers
    bias="none",  # Bias configuration for LoRA layers
    target_modules=["query", "value"]
    )

# Acquire the LoRA-adapted model
peft_model = get_peft_model(model, config)
print_trainable_parameters(peft_model)
trainable params: 2359296 || all params: 387031868 ||
    trainable%: 0.6095870120958619
# Parameters are reduced
```

From the listing 1.1 we can see that the required parameters for training are just 2.3M parameters reduced from the existing 384M parameters in which the BLIP was trained. This efficiently makes our training faster and consumes less computational resources. Here, the number of trainable parameters is 0.61% but we can make it as small as 0.01% if there are no sufficient computational resources available.

3. **Training:** We use a proprietary data loading technique for the medical image-question pairings, which efficiently batches and processes them for model training. This approach is necessary to manage the diverse and complicated nature of medical data.

   The training regimen is well thought out, using a learning rate that has been properly chosen and an optimizer such as AdamW. Using the medical VQA dataset, this optimization technique is essential for properly adjusting the model parameters. For inference tasks, optimal performance is ensured by saving the best model state, which is determined by the lowest loss.

4. **Inference and Evaluation Mechanism:** The model is used for inference on previously unseen medical image-question pairings after training. To provide pertinent responses, the refined model analyzes the questions' contextual information and the visual content of the images. The efficacy of the model in the medical VQA task is evaluated using standard metrics like open-ended accuracy, close-ended accuracy, and overall accuracy which will be further explained in the "Performance Metrics" section. These metrics assess the model's ability to accurately respond to medical questions based on the visual data.

5. **Implementation Details:** Utilizing the Huggingface Transformers module along with additional Python-based tools, the implementation provides a strong and adaptable framework for creating and implementing the model. The training process is made more transparent and manageable by using tools like tqdm for progress tracking and wandb for logging. The overall flow of our methodology can be visualized in fig 1.

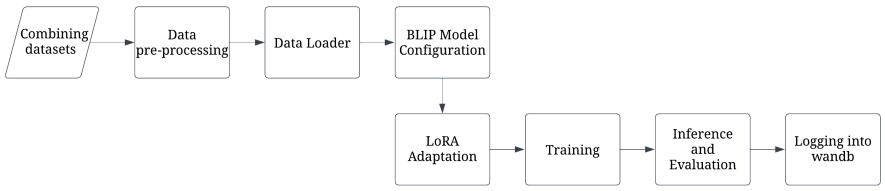

Fig. 1. Proposed methodology for LoRA-BLIP

3.5 Performance Metrics

In this study, we have used three different metrics—Open Accuracy, Closed Accuracy, and Overall Accuracy—to assess how well our optimized BLIP model performs for Medical Visual Question Answering (VQA). These measures are specifically designed to evaluate the model's performance in accurately classifying and responding to medical queries while accounting for the open-ended or closed-ended nature of the questions.

Open Accuracy : This statistic is intended to assess how well the model performs when asked open-ended questions. Open-ended questions in medical VQA refer to those whose solutions are not limited to binary choices (Answers other than Yes or No).

The formula for calculating open accuracy is:

$$\text{Open Accuracy} = \frac{\text{Number of correct OPEN predictions}}{\text{Total OPEN Ground Truth instances}} \tag{2}$$

Closed Accuracy : It evaluates how well the model performs when asked closed-ended queries, which are usually defined by binary answers (Yes or No).

The formula for calculating closed accuracy is:

$$\text{Closed Accuracy} = \frac{\text{Number of correct CLOSED predictions}}{\text{Total CLOSED Ground Truth instances}} \tag{3}$$

Overall Accuracy : It is a thorough metric that evaluates the effectiveness of the model for both open-ended and closed-ended questions. It is a micro-average accuracy that takes into account every instance, regardless of whether it is labeled as "OPEN" or "CLOSED."

The formula for calculating overall accuracy is:

$$\text{Overall Accuracy} = \frac{\text{Number of correct predictions (OPEN + CLOSED)}}{\text{Total number of instances (OPEN + CLOSED)}} \tag{4}$$

When taken as a whole, these performance indicators provide a detailed picture of the model's advantages and potential shortcomings in terms of Medical VQA. Open and Closed Accuracy offers targeted insights into particular question types, and Overall Accuracy provides an overall performance score, guaranteeing a thorough assessment of the model in an actual medical setting.

4 Result & Discussion

Our evaluation of the fine-tuned BLIP model with LoRA adaptation on various benchmark datasets for Medical Visual Question Answering (VQA) yielded insightful results. The datasets included MED-2019, RAD Dataset, and SLAKE English Dataset, each presenting unique challenges and contexts in the medical domain. The performance was assessed based on Open Accuracy, Closed Accuracy, and Overall Accuracy.

4.1 MED-2019 Dataset Results

1. **Open Accuracy:** Obtained a 78.84% accuracy rate. This shows a high ability to deal with topics that are not limited to binary choices, or open-ended questions.

2. **Closed Accuracy:** Recorded at 85%, demonstrating a strong ability to provide precise, well-defined answers to closed-ended questions.

3. **Overall Accuracy:** The model performed evenly on both kinds of questions in this dataset, as evidenced by its overall accuracy of 80.9%.

Table 3. Test accuracies on MED-2019

Metric	Value
OPEN Accuracy	78.84%
CLOSE Accuracy	85%
Overall Accuracy	80.9%

In table 3 and tab 4 we can see the accuracy metrics and the sample output visualization from wandb table respectively.

Table 4. Sample output

image ID	Question	Predicted Ans	Ground Truth
synpic22377	what imaging modality is used?	an - angiogram	an - angiogram
synpic32831	is this a noncontrast ct?	no	no
synpic59289	what type of contrast did this patient have?	iv	iv
synpic60018	is this image modality t1, t2, or flair?	t2	t2
synpic38637	is there an abnormality in the mri?	yes	yes

4.2 VQA-RAD Dataset Results

1. **Open Accuracy:** With an accuracy of 76.77%, the model's performance in open-ended questions was slightly lower. The precise imaging modalities included in the dataset or the complexity of the questions may be to blame for this.

2. **Closed Accuracy:** Obtained a score of 78.11%, indicating a respectable degree of competency on closed-ended questions.

3. **Overall Accuracy:** Although less than MED-2019, the total accuracy of 75.67% shows a respectable degree of efficacy in evaluating data and inquiries unique to radiology.

Table 5. Test accuracies on VQA-RAD

Metric	Value
OPEN Accuracy	70.77%
CLOSE Accuracy	78.11%
Overall Accuracy	73.67%

In table 5 we can see the test accuracies and table 6 shows the comparison of our model with other state-of-the-art models on VQA-RAD that currently exist. Our model is represented in bold and comparison is done based on their overall accuracy performance.

Table 6. Comparison with State-of-the-art models on VQA-RAD

Rank	Model	Open Acc.	Close Acc.	Overall
1	PMC-VQA	73.7%	86.8%	81.6%
2	MUMC	71.5%	84.2%	79.2%
3	PMC-CLIP	67.0%	84.0%	77.6%
4	M2I2	66.5%	83.5%	76.8%
5	**LoRA-BLIP**	**70.7%**	**78.1%**	**73.6%**

By this comparison, we can clearly say that our model is consistent between both open-ended and close-ended questions making it more robust and efficient.

4.3 SLAKE-English Dataset Results

1. **Open Accuracy:** Scored 77.83%, demonstrating proficiency in answering a range of open-ended medical questions.

2. **Closed Accuracy:** Scored far higher—85.82%—highlighting the model's prowess in giving accurate responses to questions with a closed-ended format.

3. **Overall Accuracy:** With an overall accuracy of 80.96%, the model demonstrated its versatility and efficacy on a wide range of SLAKE dataset question types.

Table 7. Test accuracies on SLAKE-English

Metric	Value
OPEN Accuracy	77.83%
CLOSE Accuracy	85.82%
Overall Accuracy	80.96%

In table 7 we can see the test accuracies and table 8 shows the comparison of our model with other state-of-the-art models on SLAKE-English that currently exist. Our model is represented in bold and comparison is done based on their overall accuracy performance.

Table 8. Comparison with State-of-the-art models on SLAKE-English

Rank	Model	Open Acc.	Close Acc.	Overall
1	BiomedCLIP	82.5%	89.7%	85.4%
2	MUMC	N/A	N/A	84.9%
3	CLIP-ViT w/ GPT2 (LoRA)	84.3%	82.1%	83.3%
4	BiomedGPT	N/A	N/A	81.9%
5	**LoRA-BLIP**	**77.8%**	**85.8%**	**80.9%**

The closed-ended questions are more decisive, the model consistently performs well on them across all datasets. The accuracy of open-ended questions was marginally lower, indicating a potential area for improvement as these questions frequently call for a more sophisticated understanding.

5 Conclusion

Our work refines the BLIP model improved with Low-Rank Adaptation (LoRA), which represents a breakthrough in Medical Visual Question Answering (VQA). This model's capacity for deciphering complicated medical images and providing relevant answers has been demonstrated by its adaption and evaluation of several medical datasets, including MED-2019, VQA-RAD Dataset, and SLAKE English Dataset.

In all datasets, the model performed admirably, especially when responding to closed-ended questions. This demonstrates its ability to give exact and accurate results in situations where the answers are clear-cut, which is an essential component of medical diagnostics. The model's capacity to handle a range of medical queries is demonstrated by its substantial effectiveness, albeit with slightly poorer performance on open-ended questions.

In the future, work may concentrate on improving the model's ability to answer open-ended queries and investigating how well it can be applied to a wider variety of medical datasets. Furthermore, bridging the gap between AI capabilities and clinical requirements by the integration of clinical validation and feedback into the model development process may result in Medical VQA systems that are more resilient, dependable, and clinically relevant.

To sum up, our research adds to the rapidly developing field of artificial intelligence in healthcare by demonstrating the potential of well-tuned AI models to support and enhance medical diagnosis and decision-making procedures. The encouraging outcomes of our assessments represent a step toward the development of AI-powered medical aid systems that are more clever, effective, and efficient.

References

1. Ionescu, B., Müller, H., Péteri, R., Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S. A., Kovalev, V., Kozlovski, S., Liauchuk, V., Dicente, Y., Pelka, O., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C. M., Berari, R., Tauteanu, A., Fichou, D., Brie, P., Dogariu, M., Ştefan, L. D., Constantin, M. G., Chamberlain, J., Campello, A., Clark, A., Oliver, T. A., Moustahfid, H., Popescu, A., & Deshayes-Chossart, J. (2021). Overview of the ImageCLEF 2021: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021)* (LNCS Lecture Notes in Computer Science, Springer). September 21-24, Bucharest, Romania.
2. Liu, S., Ding, H., Zhou, X.: Shengyan at vqa-med 2020: An encoder-decoder model for medical domain visual question answering task. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)

3. Verma, Harendra & Ramachandran, Sindhu. (2020). HARENDRAKV at VQA-Med 2020: Sequential VQA with Attention for Medical Visual Question Answering.
4. Chen, G., Gong, H., & Li, G. (2020). HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum* (pp. [page numbers]). CEUR Workshop Proceedings, Vol. 2696. Thessaloniki, Greece, September 22-25.
5. Jung, B., Gu, L., & Harada, T. (2020). bumjun_jung at VQA-Med 2020: VQA model based on feature extraction and multi-modal feature fusion. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum* (pp. [page numbers]). CEUR Workshop Proceedings, Vol. 2696. The University of Tokyo, Japan; RIKEN AIP, Japan. September 22-25, Thessaloniki, Greece.
6. Turner, A., & Spanier, A. B. (2019). LSTM in VQA-Med, is It Really Needed? JCE Study on the ImageCLEF 2019 Dataset. In *Conference and Labs of the Evaluation Forum*.
7. Srinivasan, K., & Noor Mohamed, S. S. (2021). SSN MLRG at VQA-MED 2021: An Approach for VQA to Solve Abnormality Related Queries using Improved Datasets.
8. Al-Sadi, A., Al-Theiabat, H., & Al-Ayyoub, M. (2020). The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG. In *Conference and Labs of the Evaluation Forum*.
9. Zhou, Y., Kang, X., & Ren, F. (2019). TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning. In *Proceedings of the ImageCLEF 2019* (pp. [page numbers]). Tokushima University, Tokushima 770-8506, JP.
10. Umada, H., & Aono, M. (2020). kdevqa at VQA-Med 2020: Focusing on GLU-based Classification. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum* (Vol. 2696, pp. [page numbers]). CEUR Workshop Proceedings.
11. Xiao, Q., Zhou, X.*, Xiao, Y., & Zhao, K. (2021). Yunnan University at VQA-Med 2021: Pretrained BioBERT for Medical Domain Visual Question Answering. Yunnan University, Kunming, China.
12. Eslami, S., de Melo, G., & Meinel, C. (2021). TeamS at VQA-Med 2021: BBN-Orchestra for Long-tailed Medical Visual Question Answering. In *Conference and Labs of the Evaluation Forum*.
13. Li, J., & Liu, S. (2021). Lijie at ImageCLEFmed VQA-Med 2021: Attention Model-based Efficient Interaction between Multimodality.
14. Hu, E. J., Shen, Y., Wallis, P., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models.