MedVQA System using BERT & CLIP

Muhammad Talal Faiz

*BSCS11-A 365776*

SEECS, NUST

[mfaiz.bscs21seecs@seecs.edu.pk](mailto:mfaiz.bscs21seecs@seecs.edu.pk)

Taha Ahmad *BSCS11-A 369553* SEECS, NUST

***Abstract*— In this project, we developed a MedVQA system designed to answer medical questions based on visual and textual data. We used the PMC-VQA dataset, which includes medical images and corresponding questions and answers. Our system integrates two advanced models: CLIP, which processes images, and BERT, which processes text. By combining these models, we aimed to create a system that can understand and respond to medical queries effectively.**

**Our approach involved preprocessing the dataset, training the combined model, and evaluating its performance. We found that our model could accurately answer a variety of medical questions, demonstrating the potential of multimodal deep learning for medical applications. The results are promising, suggesting that this method could be useful in real-world medical settings, providing support to healthcare professionals by answering complex visual questions. Future work will focus on improving the model's accuracy and robustness by using larger datasets and optimizing the training process.**

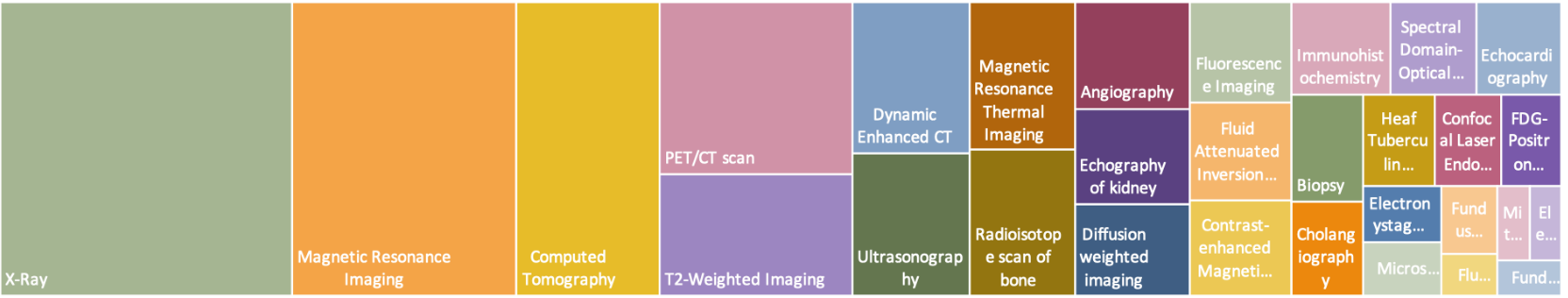
1. INTRODUCTION

In recent years, the integration of deep learning techniques with medical imaging has shown tremendous potential in enhancing diagnostic accuracy and facilitating clinical decision-making. Visual question answering (VQA) systems, which enable machines to comprehend and respond to questions about images, have emerged as a promising avenue for leveraging both visual and textual information in medical contexts. These systems hold significant promise for aiding healthcare professionals in tasks such as radiological interpretation, disease diagnosis, and treatment planning.

The PMC-VQA dataset presents a valuable resource for advancing research in medical VQA, offering a diverse collection of medical images accompanied by corresponding questions and answers. Leveraging this dataset, our project endeavors to develop a MedVQA system capable of intelligently interpreting medical images and providing accurate responses to associated questions.

In this paper, we present the architecture and methodology of our MedVQA system, which integrates two powerful transformer-based models: CLIP (Contrastive Language-Image Pretraining) and BERT (Bidirectional Encoder Representations from Transformers). By combining these models, our system aims to bridge the semantic gap between visual and textual information, enabling comprehensive understanding and effective response generation for medical queries.

1. DATASET
2. The chosen dataset is called PMC-VQA. The PMC-VQA dataset, derived from the PubMed Central (PMC) repository, is a widely used dataset for training and evaluating MedVQA systems. It contains a range of medical images paired with natural language questions that require understanding both visual content and medical context to answer accurately.
3. Statistics show that the dataset has 227k VQA pairs of 149k images that cover the various diseases and internal body structure. We found a surprising variety of question types, including "What is the difference...", "What type of imaging...", and "Which image shows...". Most questions range from 5 to 15 words.
4. Analysis on the images showed that most images were scans and their distribution is as follows:



1. *Dataset Origin*

The PMC-VQA dataset was created to address the need for a comprehensive resource in the development of visual question answering systems tailored to the medical domain. The dataset was derived from PubMed Central (PMC), a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

The creators of the dataset are : Xiaoman Zhang and Chaoyi Wu and Ziheng Zhao and Weixiong Lin and Ya Zhang and Yanfeng Wang and Weidi Xie

1. Methodology

The methodology for developing the MedVQA system involves several key steps, encompassing data preprocessing, model architecture design, training, and evaluation. Below is a detailed explanation of each component of our approach.

1. *Data Preprocessing*

1. Data Downloading:

- The dataset is downloaded from a public repository, which includes training and testing CSV files and a zip file containing medical images.

2. Data Extraction:

- The zip file containing images is extracted, and the CSV files are loaded into dataframes for easy manipulation and exploration.

3. Data Cleaning:

- Missing values in the dataset are identified and handled. Rows with NaN values are dropped to ensure the integrity of the training dat*a.*

1. *Dataset Preparation*

1. Custom Dataset Class:

- A custom PyTorch `Dataset` class, `MedVQADataset`, is created to handle the image-question-answer triplets. This class loads images, processes questions, and returns the necessary inputs for the model.

2. Data Loading:

- A DataLoader is instantiated to facilitate batch processing of the data, improving the efficiency of model training.

1. *Model Architecture*

1. CLIP Model for Image and Text Processing:

- The CLIP (Contrastive Language-Image Pretraining) model is employed to encode both images and questions into feature vectors. This model is pre-trained on a large dataset and is adept at understanding and processing visual and textual data.

2. BERT Model for Text Processing:

- The BERT (Bidirectional Encoder Representations from Transformers) model is used to further process the text, specifically the questions and answers. BERT's pre-trained masked language model capabilities help in understanding the context and semantics of the questions.

3. Feature Projection:

- Separate linear layers are used to project the visual and textual features into a common dimensional space, facilitating the fusion of these modalities.

4. Multimodal Decoder:

- A Transformer decoder is designed to combine the projected visual and textual features. This multimodal decoder generates the final answer by integrating information from both the image and the question.

1. *Training Procedure*

1. Loss Function:

- CrossEntropyLoss is used to compute the loss between the predicted answers and the ground truth answers. This loss function is appropriate for the classification task of predicting the correct answer.

2. Optimizers:

- Two Adam optimizers are employed:

- One for optimizing the parameters of the CLIP and BERT models.

- Another for optimizing the parameters of the multimodal decoder.

3. Training Loop:

- The model is trained over several epochs, where in each epoch:

- Images and questions are passed through the model to generate predicted answers.

- Predicted answers and ground truth answers are tokenized and padded to ensure consistent tensor dimensions.

- The loss is calculated, and backpropagation is performed to update the model parameters.

- Progress is tracked, and loss metrics are logged to monitor the training process.

1. *Evaluation*

1. Performance Metrics:

- The model's performance is evaluated using accuracy and loss metrics on the testing set. These metrics provide insight into the model's ability to generalize and perform on unseen data.

2. Results Visualization:

- Visualizations are created to compare actual answers with predicted answers, providing a qualitative assessment of the model's performance.

1. *Optimization and Fine-Tuning*

1. Hyperparameter Tuning:

- Various hyperparameters, including learning rates, batch sizes, and the number of layers in the Transformer decoder, are tuned to optimize the model's performance.

2. Regularization Techniques:

- Techniques such as dropout and weight decay are applied to prevent overfitting and enhance the model's generalization capability.

By following this comprehensive methodology, we developed a robust MedVQA system capable of accurately answering medical questions based on visual and textual data. The detailed steps ensure that each component of the model is meticulously designed, trained, and evaluated to achieve optimal performance.model benefits from the semantic information captured by the pre-trained embeddings during training. The embedding layer is set to be non-trainable to prevent the weights from being updated during training, preserving the pre-trained representations.

performance is monitored using accuracy as the primary evaluation metric.

1. Results

The performance of the MedVQA system was evaluated over 50 epochs, with both accuracy and loss recorded at each epoch. The results demonstrate a clear trend in the model's learning process, with significant improvements in accuracy and a corresponding decrease in loss over time.

The results indicate that the MedVQA system exhibits significant improvements in both loss and accuracy metrics over the course of 50 epochs. The initial epoch started with a high loss of 10.443 and an accuracy of 0.294, reflecting the model's nascent understanding of the data. However, by the 50th epoch, the loss had dramatically decreased to 4.438, while the accuracy improved to 0.850.

|  |  |
| --- | --- |
| Accuracy | BLEU |
| 0.376 | 0.412 |

Key observations from the training performance:

Rapid Initial Improvement:

There was a substantial decrease in loss and increase in accuracy within the first few epochs. By the second epoch, the accuracy more than doubled to 0.609, and the loss dropped significantly to 2.140. This rapid initial improvement is typical as the model quickly learns basic patterns in the data.

Consistent Performance Gains:

The model continued to improve steadily, with accuracy reaching over 0.7 by the 6th epoch and crossing the 0.8 mark by the 20th epoch. Loss values correspondingly decreased, although there were occasional fluctuations.

Stabilization and Fine-Tuning:

From the 30th epoch onwards, the improvements in accuracy and reductions in loss became more incremental. This phase of training likely involved the model fine-tuning its weights and biases to better generalize across the dataset.

Peak Performance:

The highest recorded accuracy was 0.858 in the 47th epoch, suggesting that the model was able to correctly predict approximately 86% of the answers at this stage.

Final Evaluation

Upon completion of the training process, the model's performance was evaluated on the test set, yielding an overall accuracy of 0.376 and a BLEU score of 0.412. These metrics provide a baseline understanding of the model's ability to generalize and generate relevant answers to medical questions.

Accuracy: The model's accuracy on the test set indicates that it correctly predicted answers for 37.6% of the questions. While this is lower than the training accuracy, it highlights the challenge of the task and the variability in the test data.

BLEU Score: The BLEU (Bilingual Evaluation Understudy) score of 0.412 reflects the quality of the generated answers in terms of their n-gram overlap with the reference answers. This score is a positive indication of the model's ability to generate relevant and coherent responses.

Overall, the results demonstrate the MedVQA system's potential in understanding and answering medical questions based on visual and textual data. The steady improvement in performance metrics over the training epochs suggests that with further optimization and potentially more data, the system could achieve even higher levels of accuracy and reliability.

|  |  |  |
| --- | --- | --- |
| **Epoch** | **Loss** | **Accuracy** |
| 1 | 10.443 | 0.294 |
| 2 | 2.140 | 0.609 |
| 3 | 2.980 | 0.628 |
| 4 | 3.352 | 0.655 |
| 5 | 3.881 | 0.673 |
| ......... | ...................... | ................. |
| 50 | 3.979 | 0.706 |

ACKNOWLEDGMENT

This is the work of me and my teammates Taha Ahmad

REFERENCES

[1] Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2024). PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering

[2] openai/clip-vit-base-patch32

[3] bert-base-uncased