Pretrained Encoder is all you need: An Efficient Architecture for Medical VQA

Amritpal Singh, Anisha Pal, Avinash Prabhu and Meher Shashwat Nigam
College of Computing, Georgia Institute of Technology, Atlanta, USA

***Abstract*—This course project focuses on medical visual question answering (VQA) using deep learning techniques. Specifically, we propose a simple but highly accurate VQA model that combines visual and language input encodings before passing them to a classifier. We conduct experiments using various vision and language encoders and demonstrate that the choice of encoder significantly impacts the model's accuracy. Our results show that using novel vision encoders can achieve comparable accuracy to state-of-the-art models. Additionally, we provide a flexible and easy-to-use framework for training and testing VQA models with different encoders on medical VQA datasets. Overall, our work highlights the importance of encoder selection and offers a practical solution for VQA in the medical domain. Link to our code and dataset - github.com/Anipal/NLP-CS7650-Project**

I. Introduction

Medical Visual Question Answering (VQA) is a research area that aims to answer natural language questions about medical images. The ability to automatically interpret and understand medical images using natural language can smoothen patient interaction with healthcare systems. When resources are limited, medical VQA can offer a "second opinion" to radiologists on their image analysis, and patients can obtain basic information about medical images without seeing a doctor.

Medical VQA faces unique challenges compared to normal VQA due to diversity of questions that can be asked, complex/nuanced nature of medical images, which often contain a wide range of visual features like abnormalities, lesions, and anatomical structures that require specialized knowledge to interpret. Medical VQA also requires domain-specific vocabulary that is not commonly used outside the medical field, such as terms like "hypodensity", "hematoma," and "metastasis". Furthermore, medical VQA models must be able to interpret the visual features in the context of medical diagnosis and treatment, requiring a more nuanced understanding of the relationship between the visual features and the underlying pathology.

These unique challenges make developing accurate and reliable medical VQA models harder than the normal VQA domain. Medical images, particularly in radiology, require the interpretation of clinical history or multiple images from different views/planes to support medical professionals' decision-making. The lack of this additional multimodal/multi-view information in the inference procedure adds to the difficulty of the task. The VQA-RAD research project has investigated the natural questions that arise in clinical conversations and has categorized them into different types such as modality, plane, organ system, abnormality, object/condition presence, positional reasoning, color, size, attribute other, counting, and others. In this work, we frame medical VQA as a classification problem. We experiment with different kinds of image and text encoders, aggregation methods and final classifiers.

II. Related Work

A. Common methods

The "joint embedding" method is the most common approach, also proposed as the baseline method for the VQA v1 [14] dataset. The architecture usually comprises an image encoder, a question encoder, features fusing algorithm, and an answering component according to the task requirement. Many new methods use attention mechanisms, like [16] which use a question-conditioned reasoning module to guide the importance selection over multimodal fusion features to obtain SOTA results on VQA-RAD. Other works utilizing attention use Stacked Attention Networks(SAN) [92], Bilinear Attention Networks (BAN), and Hierarchical Question-Image Co-Attention (HieCoAtt), etc [15].

B. Medical VQA datasets

There are many publicly-available medical VQA datasets up to date [15]: VQA-MED-2018, VQA-RAD, VQA-MED-2019, RadVisDial, PathVQA, VQA-MED-2020, SLAKE, and VQA-MED-2021 (in chronological order). The question categories include color, modality, abnormality, organ, position, and more.

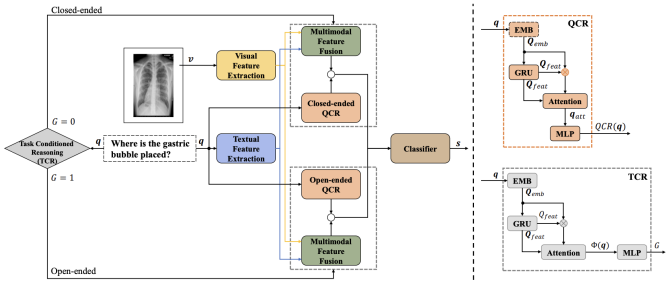


Fig. 1: Baseline Architecture

III. Methods

A. Dataset

For this work, we use the VQA-RAD dataset [1]. VQA-RAD contains 315 radiological images sampled from an online

radiology database. Images were sampled based on the region of the scan (head, chest, and abdomen), with a balanced representation from each region. Questions (Open & closed-ended) and answers were generated for images by medical trainees and verified by experts. Fig. 2 (a) and (b) shows an example of closed-type, open-ended question. These questions are paraphrased, leading to a final pool of 3515 questions. Fig. 2 (c) shows a paraphrased version of the question (b). Fig. 3 shows dataset distribution based on the question type, answer type, phrase type, and organ in the image.

Open-ended questions allow for a wide range of possible answers, which are not restricted to a particular set of options. This requires understanding the content of the image and the language of the question to generate a relevant response. Generative approaches perform better on open-ended questions. On the other hand, closed-ended questions typically have a fixed set of answer choices(e.g. Yes, No or a list of possible options), and the model must select the correct answer from the available options. These questions are typically easier to answer since the possible answers are restricted. This is generally framed as a classification problem.

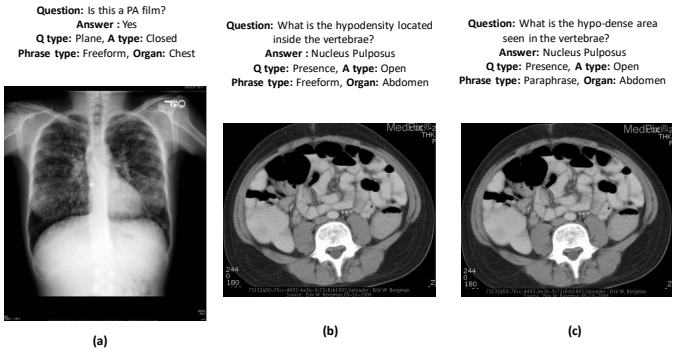


Fig. 2: VQA-RAD dataset. (a) Closed type question for Chest region (b) Open type question for Abdomen scan (c) Paraphrased question for the same scan

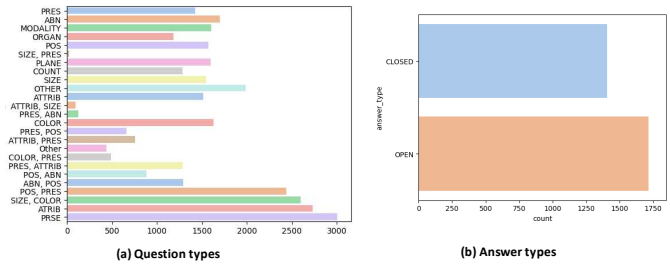


Fig. 3: Distribution of questions and answers in VQA-RAD dataset

For the evaluation of architecture, we use accuracy measures on open-ended type, close-ended type, and the entire test dataset.

B. Image encoders

1) EfficientNet: Previous methods for scaling CNNs (such as increasing depth, width, or resolution) often led to diminishing returns in terms of accuracy and efficiency. EfficientNet [6] was proposed a as new method that balances all three dimensions using a compound coefficient that can be easily adjusted based on the available resources. Neural architecture search was used to design a new family of models called EfficientNets, which achieve state-of-the-art accuracy and efficiency on various datasets while being smaller and faster than previous CNNs. The architecture of EfficientNet consists of a stem, multiple blocks with different depths and widths, and a head with global average pooling and fully connected layers.

2) Swin Transformer: The Swin Transformer [5] was designed to create a general-purpose transformer-based architecture to address the challenges in adapting the transformer from language to vision. To address these challenges, Swin Transformer uses a hierarchical architecture that starts with small-sized patches and gradually merges neighboring patches in deeper Transformer layers. Additionally, Swin Transformer uses shifted windows to reduce the computational complexity of self-attention on high-resolution images. This results in a linear computational complexity to image size, making it more efficient than traditional Transformers for computer vision tasks.

3) ConvNext: The ConvNext model [4] was designed to explore the space of convolutional neural networks and challenge the widely held belief that Vision Transformers are superior to CNNs. The architecture of ConvNeXt combines parallel pathways of convolutions with different kernel sizes and depths to capture multi-scale features effectively. It achieves state-of-the-art performance on multiple computer vision benchmarks while retaining the simplicity and efficiency of standard CNNs.

C. Language Encoders

The language encoders used in this work can be divided into two categories - BERT and RoBERTa.

1) BERT: BERT (Bidirectional Encoder Representations from Transformers) [7] is a language representation model designed to pretrain deep bidirectional representations from unlabeled text by joint conditioning on both left and right context in all layers. Its architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. [8]. The motivation behind BERT is to achieve state-of-the-art performance on a wide range of language tasks by leveraging large amounts of unlabeled data. We used the following language encoders based on BERT -

- **BERT**: The original BERT model described above.
- **BioClinical BERT**: This model [10] is initialized with the BERT base and finetuned with all notes from MIMIC III, a database containing electronic health records from ICU patients.
- **SciBERT**: This model [11] is initialized with the BERT base and finetuned on scientific text taken from *Semantic Scholar*.
- **BlueBERT**: This model is initialized with the BERT base and finetuned on *PubMed* abstracts.

2) RoBERTa: The RoBERTa [9] model was proposed after investigating the impact of hyperparameters and training data size on the performance of BERT pretraining. This approach involves several modifications, including larger batch sizes, longer training times, longer sequences, and dynamically changing the masking pattern applied to the training data. This model was trained on a large new dataset (CC-NEWS) to better control for training set size effects. RoBERTa achieved state-of-the-art results on several benchmark datasets, including GLUE, RACE, and SQuAD. Overall, the RoBERTa model demonstrates the importance of carefully considering design choices in language model pretraining and highlights the potential for further improvements in this field.
We used the following language encoders based on RoBERTa-

- **BioMed-RoBERTa**: This model [12] is initialized with the RoBERTa base and finetuned on a scientific text taken from *Semantic Scholar*.

3) CLIP: : CLIP [13] is used to encoder *both* images and text. Given image and text descriptions, the CLIP model can predict the text description for the image without optimizing for a particular task.

D. Classifiers

In general, VQA is considered a classification task as the goal is to predict a class label from a pre-defined set of classes, instead of generating a new answer from scratch. In adherence to this, we treat our task as a classification task. Specifically, the VQA-RAD dataset can be grouped into 458 classes. The model's task is to predict the appropriate class given an image and question pair.
1) Linear layers: The classifier part of the model, which comes after combining the image and text feature consists of 3 linear layers, with a dropout layer (p=0.2) after the first layer.
2) Linear layers with Attention: The classifier part of the model, with attention modules. Combined image and text features are passed through 3 attention modules, followed by 3 linear layers, with a dropout layer (p=0.2).

E. Final Architecture

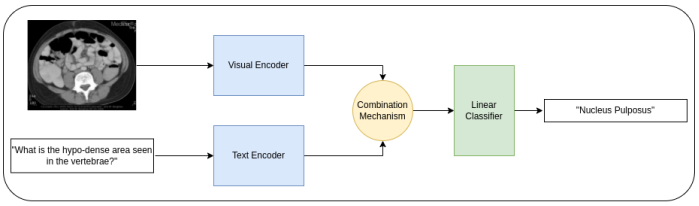


Fig. 4: Architecture Used

The architecture used in the experiments is a straightforward design, as illustrated in Fig. 4. It incorporates two parallel branches, one for extracting image embeddings and the other for extracting text embeddings, which are subsequently fed into a classifier model to generate a prediction. The beauty of this architecture lies in its simplicity, which allows for easy swapping of different pretrained encoders and classifiers. This facilitates the comparison and analysis of the contributions of individual components toward achieving the final results. The image and text embeddings are generated offline and fed into the classifier model. The classifier is the only component of the architecture which is trained on the VQA-RAD dataset.

F. Novelty

The novelty of our method lies in how simple it is. To the best of our knowledge, no prior work in the field of medical VQA has proposed a method as straightforward as the one presented here. Despite the model's simplicity, the achieved results are competitive with state-of-the-art approaches, which prompts the question of whether complex models, which have been commonly proposed in the past, are truly necessary for medical VQA tasks. The promising outcomes of our proposed method highlight the potential for simpler models to perform comparably to more complex ones in this domain.

IV. Experiments

A. Baseline

We considered the work by Zhan et al. as our baseline. To the best of our knowledge, they report SOTA results on the dataset we have considered- VQA-RAD. They utilize a question-conditioned reasoning module to guide the importance selection over multimodal fusion features. We used their publicly available code and were able to reproduce the results presented in the paper.

B. Model setup and experiments

The model was trained using Cross Entropy loss and Adam optimizer with a learning rate 0f 0.001. The numbers reported in Table I are on the test dataset of 408 datapoints containing 128 open-ended types and 280 close-ended types belonging to a total of 83 different classes.
Experiments were conducted to evaluate the importance of the following components as illustrated in Table I:

- Text Encoders
- Image Encoders
- Aggregation method
- Classifier

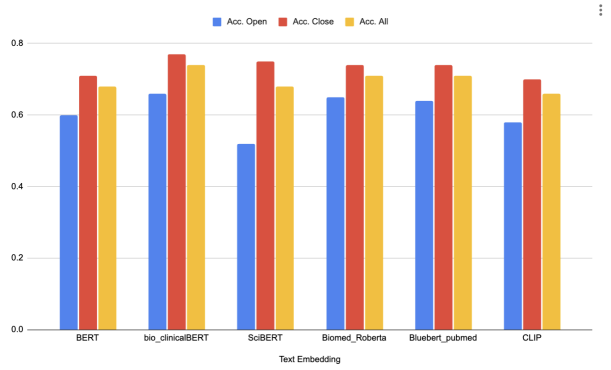
V. Results and Discussion

We summarize our key findings below :

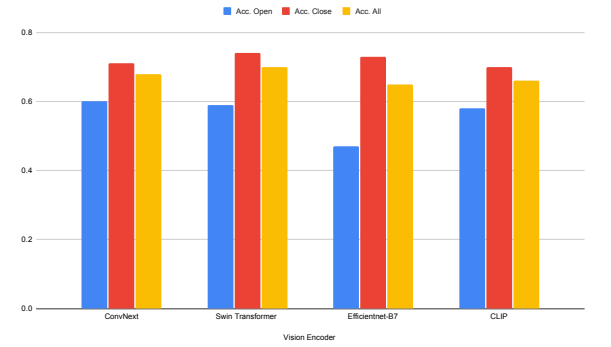
▷ **Strong pretrained encoders provide the largest gains**
As shown in Table I, using a powerful pre-trained encoder such as BERT for text and ConvNeXt for images can yield an accuracy of approximately 94.%, which is comparable to the baseline [16]. Stronger text encoders are optimized to capture more nuanced relationships between words, resulting in more meaningful and informative embeddings that can be more effectively decoded by the classifier. Similarly, stronger image encoders can generate more visually meaningful features that highlight the important and distinguishing aspects of an image. As seen from our results highly informative embeddings can generate equally good results with a very lightweight

Text Encoder	Image Encoder	Aggregation	Classifier	Acc. Open	Acc. Close	Acc. All
GLoVe*	MEVF*	(QCR + TCR $\rightarrow$ MUL)*	MLP*	0.6*	0.79*	0.72*
BERT	ConvNeXt	MUL	MLP	0.6	0.71	0.68
BERT	ConvNeXt	ADD	MLP	0.52	0.7	0.64
BERT	ConvNeXt	MUL	Attention	0.57	0.71	0.67
BioClinicalBERT	ConvNeXt	MUL	MLP	0.66	0.77	0.74
SciBERT	ConvNeXt	MUL	MLP	0.52	0.75	0.68
BiomedRoBERTa	ConvNeXt	MUL	MLP	0.65	0.74	0.71
BlueBERT	ConvNeXt	MUL	MLP	0.64	0.74	0.71
Swin Transformer	BERT	MUL	MLP	0.59	0.74	0.7
Efficientnet-B7	BERT	MUL	MLP	0.47	0.73	0.65
CLIP	CLIP	MUL	MLP	0.58	0.7	0.66

TABLE I: The accuracy of open-ended, close-ended, and all answers is compared for each combination of text and image encoder, aggregation method, and classifier used in the experiments. The symbol (*) corresponds to the baseline architecture mentioned in [16].



(a)



(b)

Fig. 5: Comparing the performance of (a) different text encoders in conjunction with the ConvNeXt image encoder, (b) different image encoders in conjunction with the BERT text encoder. For CLIP same encoder is used for both image and text

classifier hence resulting in an extremely computationally efficient and interpretable model.

▷ **Context is important**

The results presented in Fig. 5 (a) demonstrate that BioClinicalBERT exhibits the best performance, with comparable results from BioMed-RoBERTa and BlueBERT. Each of these three text encoders has been trained on a large corpus of relevant medical text to capture relationships between medical terms and conditions and the ways in which medical professionals communicate about symptoms and treatments.

Encoders trained on medical data are therefore better suited to capturing the nuances of medical language and terminology, leading to improved performance. This is further supported by the findings presented in Fig. 5 (b), where the image encoders pre-trained on generic, non-medical data are compared while keeping the text encoder fixed to BERT. In this scenario, a sharp decrease in performance is observed compared to Fig. 5 (a), due to a lack of relevant context.

▷ **Classification of open-ended answers is a difficult task**

From Table I it is clearly reflected that the difference in performance for open-ended vs close-ended answers is approximately $\geq 10\%$. And this behavior is constant across all the models proving that classifying open-ended answers is a difficult task. Unlike close-ended answers, open-ended answers can vary in the number of reasoning steps required to generate the required answer. Additionally, open-ended answers require deeper analysis to identify the key themes or concepts and to infer the sentiment or opinion expressed in the response making it a more difficult classification problem. Using a generative model can help tackle these questions better.

VI. Conclusion

In this study, we introduce a highly lightweight architecture with a user-friendly plug-and-play design that serves as an effective platform for examining the significance of different components in Medical VQA model design. Our experiments highlight the critical role of leveraging the power of encoders trained on significantly large corpora of contextually relevant data.

Going forward, it would be crucial to explore stronger classifiers and consider a generative approach to addressing this problem, rather than a classification-based one, to generate more meaningful outcomes.

REFERENCES

[1] Lau, J., Gayen, S., Ben Abacha, A. et al. A dataset of clinically generated visual questions and answers about radiology images. Sci Data 5, 180251 (2018). https://doi.org/10.1038/sdata.2018.251

[2] Nguyen, B.D., Do, TT., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D. (2019). Overcoming Data Limitation in Medical Visual Question Answering. In: , et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science(), vol 11767. Springer, Cham. https://doi.org/10.1007/978-3-030-32251-9_57

[3] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. 2021. Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering. In Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21). Association for Computing Machinery, New York, NY, USA, 456–460. https://doi.org/10.1145/3460426.3463584

[4] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., amp; Xie, S. (2022). A convnet for the 2020s. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01167

[5] 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. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992-10002.

[6] : Tan, M. amp; Le, Q.. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ¡i¿Proceedings of the 36th International Conference on Machine Learning¡/i¿, in ¡i¿Proceedings of Machine Learning Research¡/i¿ 97:6105-6114 Available from https://proceedings.mlr.press/v97/tan19a.html.

[7] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

[8] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is All you Need. ArXiv, abs/1706.03762.

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.

[10] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W., Jin, D., Naumann, T., McDermott, M.B. (2019). Publicly Available Clinical BERT Embeddings. ArXiv, abs/1904.03323.

[11] Beltagy, I., Lo, K., Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Conference on Empirical Methods in Natural Language Processing.

[12] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ArXiv, abs/2004.10964.

[13] Wang, Z., Codella, N.C., Chen, Y., Zhou, L., Yang, J., Dai, X., Xiao, B., You, H., Chang, S., Yuan, L. (2022). CLIP-TD: CLIP Targeted Distillation for Vision-Language Tasks. ArXiv, abs/2201.05729.

[14] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., Parikh, D., 2015. VQA: Visual question answering, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 2425–2433

[15] Lin Z, Zhang D, Tac Q, Shi D, Haffari G, Wu Q, He M, Ge Z. Medical visual question answering: A survey. arXiv preprint arXiv:2111.10056, 2021.

[16] Zhan Li-Ming, Liu Bo, Fan Lu, Chen Jiaxin, and Wu Xiao-Ming. 2020. Medical visual question answering via conditional reasoning. In Proceedings of the 28th ACM International Conference on Multimedia. 2345–2354

VII. CONTRIBUTIONS

- Amrit - coding up the deep learning model, training and testing, visualizations, dataset literature survey, Introduction and Dataset writing
- Anisha- coding up the deep learning model, training and testing, methods literature survey, Experiments, Results Discussions and Conclusion writing.
- Avinash - coding up the deep learning model, training and testing, methods literature survey, Methods writing
- Meher - coding up the deep learning model, baseline training and testing, methods literature survey, Related works, and abstract writing