Designing and Training a Fully Attentive Multimodal Transformer Network for Medical Visual Question Answering Task

Md Mesbahur Rahman

Major
Master of Computer Science (Online), Department of Computer Science
The University of Texas at Austin

ABSTRACT

Medical Question Answering is a very important and impactful application of Multi-modal learning. It can contribute to the interpretability of machine learning model in medical applications, reduce workload of medical professional, and can be a part of fully automated healthcare system. In this project, we have done a background research on the state of the art of Medical Visual Question Answering research. Based on some latest well performing paper, we propose our own fully attention based Transformer only network for solving the medical visual question answering task by treating a multi-class classification problem. We also present some analysis on hyperparameter tuning of the model, compare its performance with models from some other notable papers and suggest some future improvements of our model.

1. INTRODUCTION

Medical visual question answering opens a new way to interact between the AI models and the physicians on the diagnosis of the disease. It is also a significant application of multi-modal learning, which can significantly, which aims to increase the learning capacity and applicability of AI models by fusing two or more types of data source or modality. In visual question answering, two types of information namely, image and text are combined and then fed into AI models to generate an answer for an input query image and textual question pair. Medical Image question answering are special in a sense that the characteristics of medical images are different from regular RGB images and also the expected accuracy of explanation are more strict due to the criticality of their medical applications. Medical VQA system can significantly contribute to the efficiency improvement of the medical professionals, by giving a second opinion and support of confidence in medical diagnosis. They can also be part of a much larger medical knowledge base. They can be part of a fully automated medical diagnosis system where an expert physicians are not simply available. In this project, we explored the realm of the medical visual question answering and state-of-the-art research in this special field. We train a transformer-only network for building a medical question answering system, and then we do some analysis on the trend observable in the result of our training.

2. RESEARCH BACKGROUND or LITERATURE REVIEW

The first completion on Medical Visual Question answering was held in 2018 called ImageCLEF [1]. According to a survey done by Lin et el [2] eight open source datasets which are focussed on Medical VQA. They are namely VQA-Med-2018 [1], VQA-RAD [3], VQA-Med-2019 [4], RadVisDial [5], PathVQA [6], VQA-Med-2020 [7], SLAKE [8], VQA-Med-2021 [9]. These datasets differ mainly in domain of medical images (e.g., X-ray, CTs, MRI and Pathology) and also respective organ systems. They also differ in modality of tasks such as VQA, segmentation, Due to the expensive nature of medical image annotation, the number of images in Medical VQA datasets are comparatively lower than the more general purpose VQA datasets. In short, the development of Medical VQA datasets are in still very early age and their data subjects are very limited.

Regarding the method for solving the Medical VQA problem, there is an obvious common pattern called joint embedding [10]. This particular approach is composed of an image encoder, a question encoder, a feature fusion component and a task specific head. The image feature extractor can be any well suited image backbone network like VGG [11], ResNet [12] or even a vision specific Transformer like ViT [13] and Swin-Transformer [14]. For question encoding, one can choose any efficient language model like LSTM [15], GRU [16] or Transformer [17].

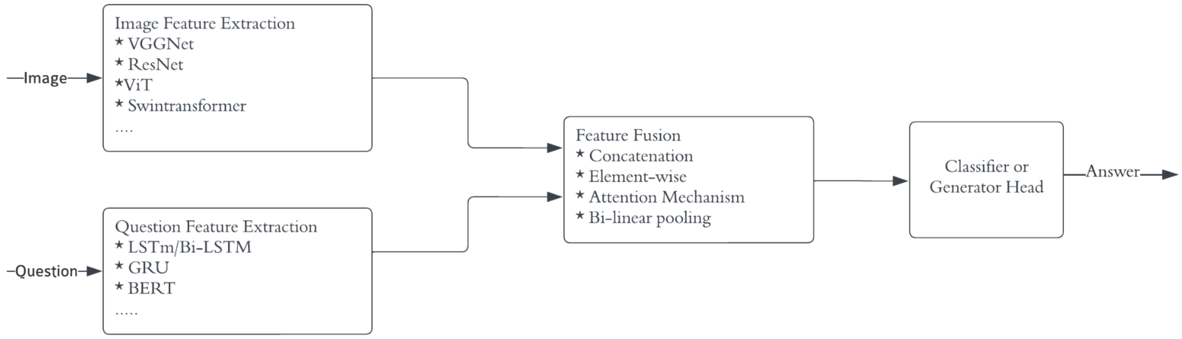


Figure 1. Common system design for Medical-VQA model

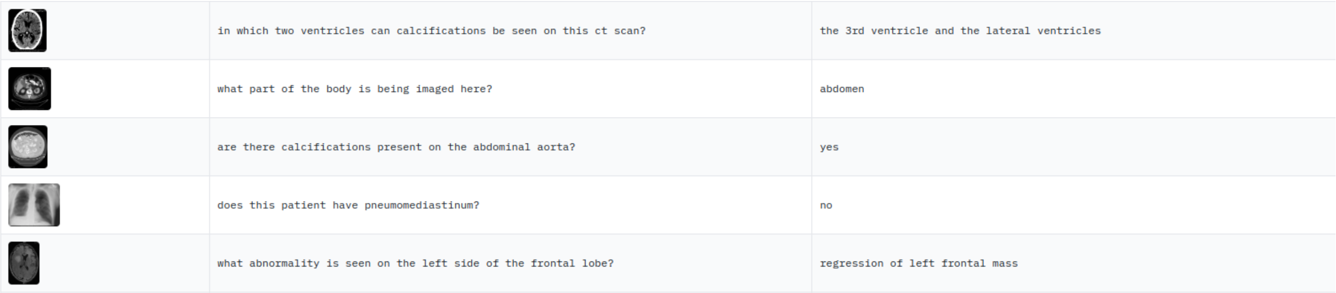
The next and arguably the most important component of a Medical VQA system is the Feature fusion block. In order to extract information effectively from both the Image and Question, and, we have to combine these two features and correctly formulate the relationship between the two feature set. There are several feature fusion techniques can be seen in the literature. One prominent techniques is to concatenate and apply some form of linear transformation like summation, multiplication etc. But these operations are often expensive that is why some researchers often uses direct concatenation, but the result is often mediocre. And some reachers use some efficient pooling techniques to first concatenate the image and question features to higher dimension and then use some coevolving operation to reduce their dimensionality. Some notable pooling techniques are Multi-modal Compact Bilinear pooling [18], Multi-modal Factorized High-Order (MFH) pooling [19], etc. Another main school of feature fusion is to use attention mechanism to feature fusion. Some notable attention mechanism for feature fusion are Stacked Attention Networks (SAN) [20], Hierarchical Question-Image Co-Attention (HieCoAtt) [21], Bilinear Attention Networks (BAN) [22] etc. Although multi-head attention mechanisms namely, Transformer [17] etc. are popular in general LLMs, but they rarely used for Medical-VQA.

The final component is the task specific head, which are mainly of two types depending on the two schools of Medical-VQA models, one of classifier type and the other is of generative nature. In both cases, the generally several layers deep fully connected network is used as feature extractor for the classification or generation purpose. The classification approach works well for a small search space of answers, and the generation approach works well for open-ended cases. Some papers ([23], [24], [25], [26] etc.) use a switching strategy between the two approaches.

There are mainly two types of metrics to evaluate on Medical VQA tasks. One is classification based metrics (e.g., accuracy, f1-score etc.) and another is language based metrics (e.g., BLEU [27] etc.) which focuses on response sentence evaluation for image captioning, report generation tasks [28].

3. MATERIALS / DATA / SOURCES

For training and evaluating our Medical VQA algorithm, we chose VQA-RAD [3] dataset for our experiment. It has both closed and open-ended questions, and it is the only dataset that contains natural question and categories distribution from medical students. It has 315 images and, 2248 question pairs [29]. The image is collected from MedPix and it images are balanced across three human organs (namely head, chest, and abdomen). It is publicly available on Open Science Framework and HuggingFace [29]**.** We also used the train-test split provided by the authors [3] of the dataset.



	in which two ventricles can calcifications be seen on this ct scan?	the 3rd ventricle and the lateral ventricles
	what part of the body is being imaged here?	abdomen
	are there calcifications present on the abdominal aorta?	yes
	does this patient have pneumomediastinum?	no
	what abnormality is seen on the left side of the frontal lobe?	regression of left frontal mass

Figure 2. Sample Image-question-answer triplet [29]

4. METHODS

In our experiment, we tried to build a transformer only network for addressing Medical-VQA challenge. Because, Transformers are very good at paying attention to different part of the feature set and give appropriate importance to what is necessary to answer the question about the image. We primarily took inspiration from two very recent papers on Medical-VQA (e.g. Q2ATransformer [30] and Multi-modal Pre-training [31]. Similar to figure 1. We chose our own algorithms for each component of the general Medical VQA system. We have described the components in detail below.

4.1. **Image Encoder**

The main function of Image Encoder is to extract import information for visual comprehension and question comprehension from the input image. We can pick any visual feature extractor as the image encoder algorithm. It could be a CNN based network like VGG [11], ResNet [12] or a vision Transformer model like ViT [13] and Swin-Transformer [14]. For the image encoder component of our Medical-VQA system, we picked pre-trained SwinTransformer because of three reasons. First, for medical image feature extraction Swin-ransformer is more suitable compared to CNN based network because Swin-transformer computes through cross-window connection and hence able to extract relations and information between distant grid and very fine-grained feature. Second, similar to many hierarchical CNN vision model, Swin-transformer network can exploit image feature at different scale and increment in image size cause the computation to grow only linearly. Third, similar to VGG [11], ResNet [12] Swin-Transformer was also trained on very large image set and are very effective and efficient to extract useful image feature. Swin-Transformer divides the input image into several non-crossing visual tokens and fed into attention network to extract visual features. So our pre-trained Swin-Transformer image encoder takes a input image and gives an output of $V = [v_1, v_2, ..., v_N] \in R^{N\times d}$. Where N is the number of visual tokens and d is the number of embedding to be as input to the feature fusion component.

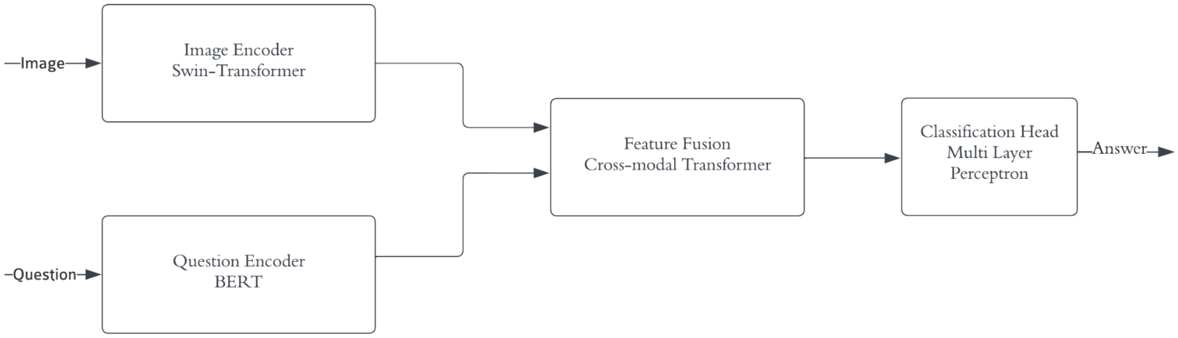


Figure 3. Our proposed system architecture for Medical Question Answering

4.2. Question Encoder

For question encoder, we chose a pre-trained language encoder like BERT although the paper “Multi-modal Pre-training” [31] utilize pre-trained tokenizer named WordPiece [32] mainly because BERT is very good at question comprehension by exploiting bidirectional attention. Since we built this system primarily for solving the visual question answering task as a classification problem, bidirectional attention is not an issue. But it would be an issue if we followed a generative approach, in that we would have to utilize a seq2seq attention mask. BERT extracts question embedding is by $\mathrm{BERT}(Q_e) = F_q \in R^{M\times d}$ where $Q_{e,}$ M, d are respectively the input question, number of text embedding feature and dimension of the output embedding.

4.3. Feature Fusion Component

For feature fusion component instead of using the CMAN module proposed by Q2ATransformer [30], we utilized the Cross-modal Transformer encoder proposed by the Multi-modal Pre-training [31], which consists of a multi-head self attention layers. The different heads of the cross-modal transformer can attend to different part of input embeddings and extract different types of cross-modal relationships between the image and textual feature sets. Inside the Cross-modal Transformer, we mainly incorporated bidirectional attention mask so that each token can attend to other tokens on its both sides.

4.4. Task Specific Head

We generated an answer for an image-question pair by treating the question-answering task as a classification, where we selected the most probable answer from a pool of possible answers. To be exact there are 458 unique answers in the train dataset so we consider the question answering task as 458 class classification problem. A classifier consists of multi-layer perceptron is attached on top of the token obtained from the cross-modal Transformer. During training, we optimized **Categorical Cross-entropy** [33] loss implementation from PyTorch and during inference we take the *argmax* of *softmax* [34] probability of the logits predicted by the model to predict the most probable class of answer.

Note: For all the Transformer implementation (e.g. Swin-Transformer [14] and BERT we utilized the implementation provided by the **transformers** library from HuggingFace.

5. RESULTS

There are three standard metrics for evaluating models' performance on Medical VQA tasks, specially VQA-RAD dataset. Overall accuracy, open-ended accuracy and close-ended accuracy. We also evaluated our model on the test-set of VQA-RAD using these three metric as well.

Overall Accuracy = $\frac{total\ no.\ of\ correct\ answers}{total\ no.\ of\ questions}$

Open-ended Accuracy = $\frac{total\ no.\ of\ correct\ answers\ for\ open-ended\ questions}{total\ no.\ of\ \ open-ended\ question}$

Close-ended Accuracy = $\frac{total\ no.\ of\ correct\ answers\ for\ open-ended\ questions}{total\ no.\ of\ open-ended\ questions}$

It is to be noted that for even for open-ended questions, the predicted answer has to be in exact match with the ground truth answer in order to consider that prediction as correct.

We experimented with different learning rate parameter to see its effect on the accuracy of our proposed model on test-set performance of VQA-RAD. And below we plotted three different plots for open-ended, close-ended and overall accuracy with each plot containing three learning curves for learning rate = 2e-5, 2.5e-5 and 3e-5.

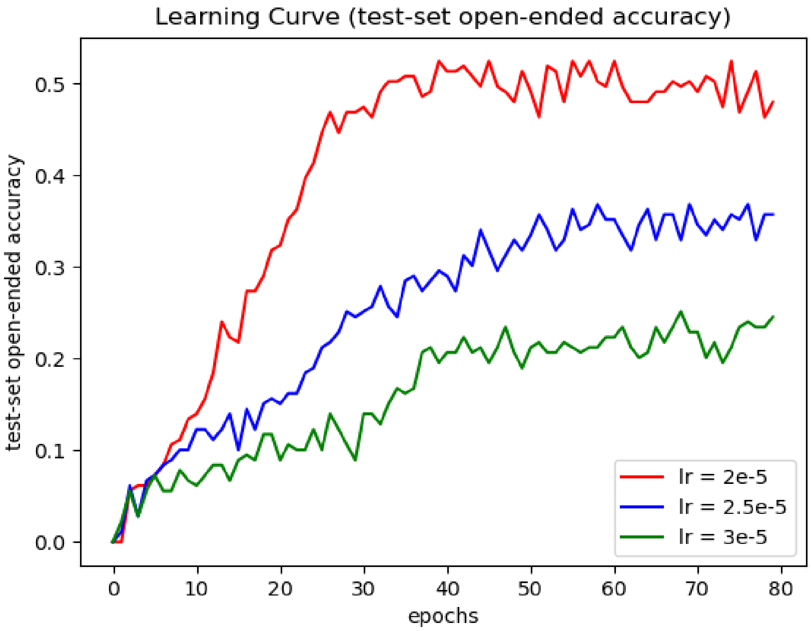


Figure 4. Learning Curve in terms of test-set open-ended accuracy

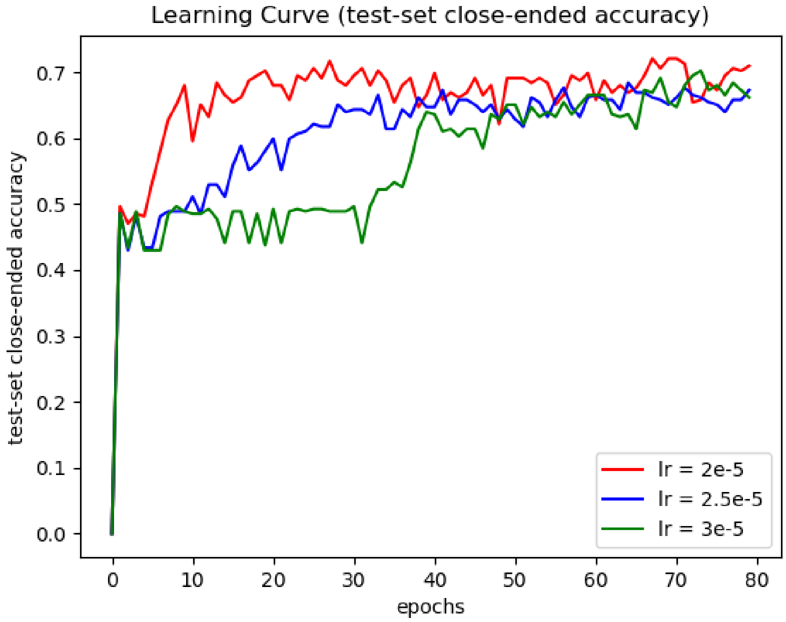


Figure 5. Learning Curve in terms of test-set close-ended accuracy

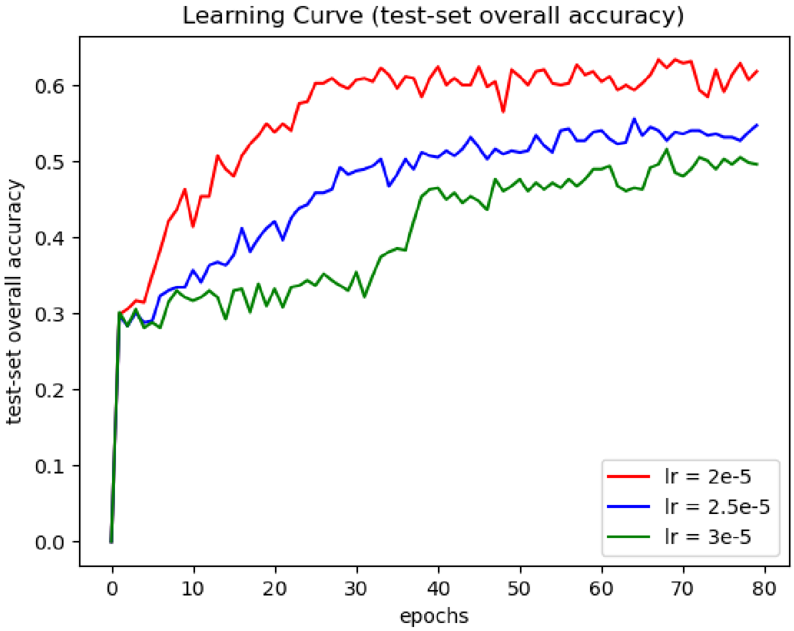


Figure 6. Learning Curve in terms of test-set overall accuracy

We can see learning rate variation does not have much impact on the close-ended accuracy, but it greatly affects the open-ended accuracy and hence the overall accuracy. Reducing learning rate largely improve the open-ended accuracy as can be seen from the figure 4.

The next experiment we did with the was with the activation function used at the end of encoders and decoder (namely Swin-Transformer and BERT). We experimented between ReLU [35] and GELU [36] compared open-ended, close-ended and overall accuracy between the two in figure 7.

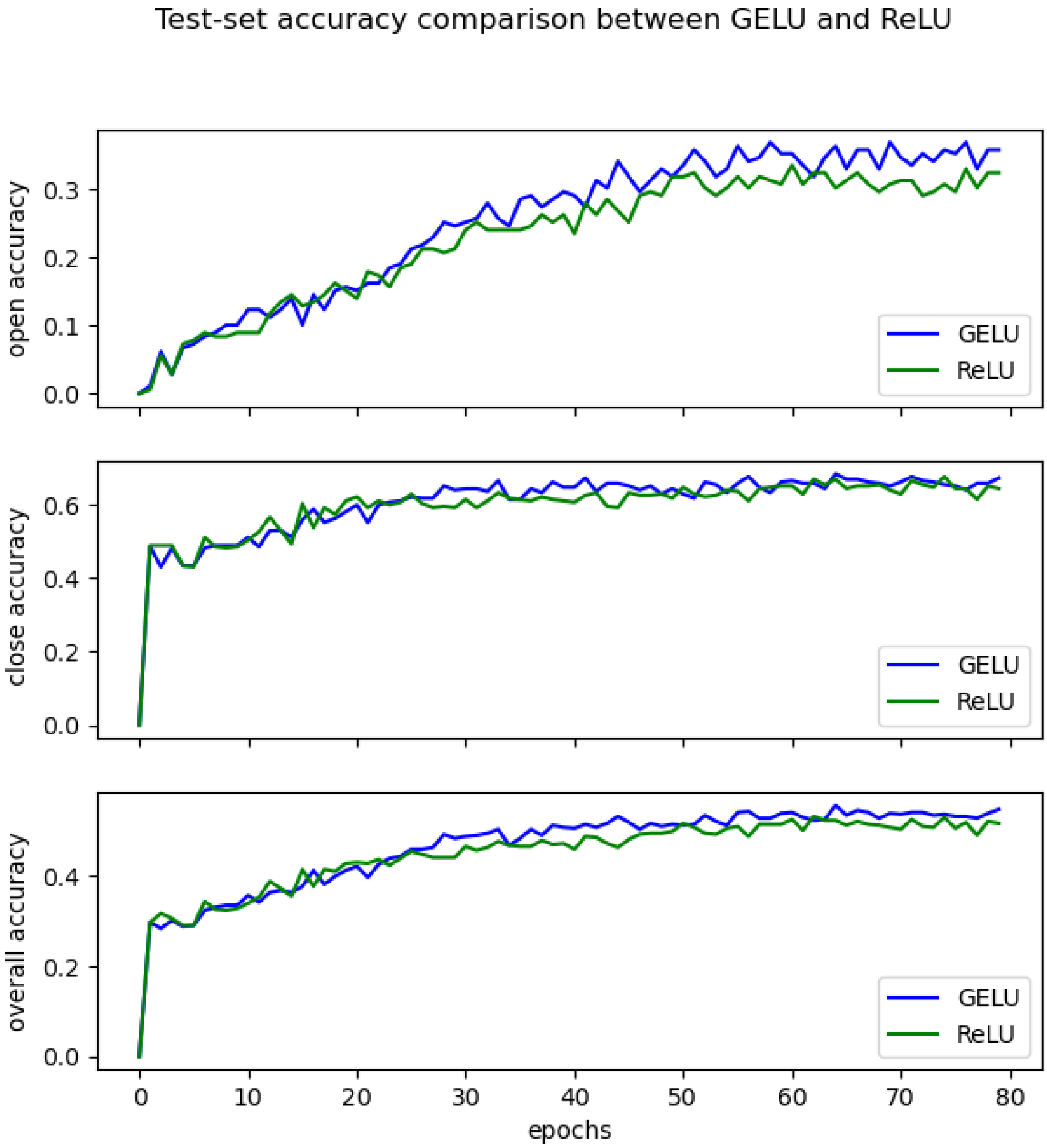


Figure 7. Learning Curve comparison between GELU and ReLU

As can be seen from figure 6. GELU always performs better in comparison to ReLU in all three metrics, and the difference in performance is more noticeable in open-ended accuracy.

Our final experiment was with the attention mask use in the feature fusion component, Cross-modal transformer. We experimented with both bidirectional attention mask and seq2seq (unidirectional) attention mask and visualized the result in figure 8.

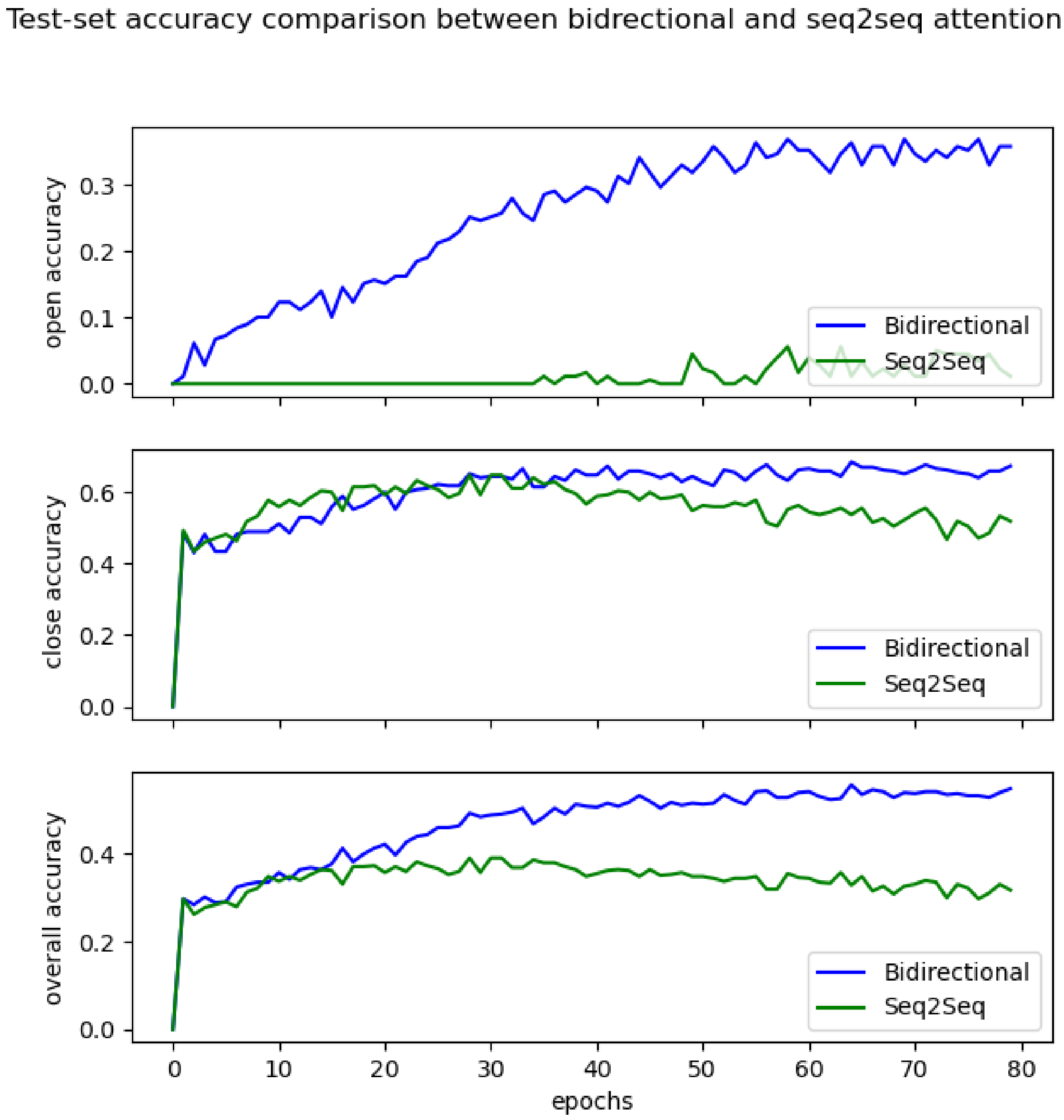
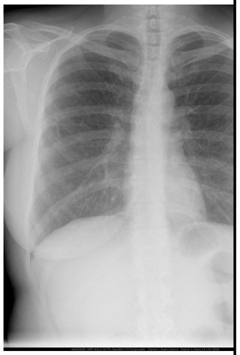
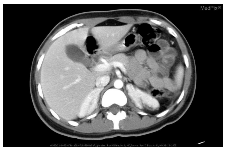
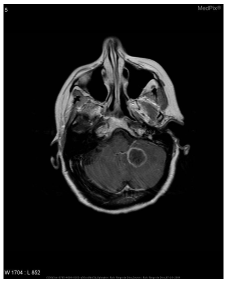
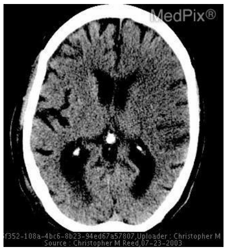


Figure 8. Learning Curve comparison between bidirectional and seq2seq attention mask

It can be seen from the above figure that bidirectional attention mask always performs better that seq2seq mask in our implementation for visual question answering system. The difference is more prominent in open-ended accuracy. This might be due to the fact that we calculated our loss in multi-class classification fashion, where proper generative model require training with sequential cumulative loss calculation.

We have also some sample prediction results obtained from our best model here.

Table 1. Sample prediction Image Question Answer Triplets

Image	Question	Predicted Answer	Correct Answer	Answer Type
	does this represent adequate inspiratory effort?	yes	yes	Closed
	is the gallbladder present?	no	yes	Closed
	what are the black areas at the top of the image?	maxillary sinuses	maxillary sinuses	Open
 MedPix® f352-108a-4bc6-8b23-94ed67a57807,Uploader : Christopher M Source : Christopher M Reed,07-23-2003	in which two ventricles can calcifications be seen on this ct scan?	non contrast ct	left mca	Open

In the following table (table 2.) we have compared the test accuracy of our best model (learning rate = 2e-5, epoch=80) with some notable papers on Medial VQA task along with our model inspiring papers Q2ATransformer [30] and Multi-modal Pre-training [31].

Table 2. Test accuracy comparison

Model	Open-ended Accuracy	Close-ended Accuracy	Overall Accuracy
Q2ATransformer [30]	79.19	81.2	80.48
Multi-modal Pre-training [31]	72.1	60.9	79.4
MMBERT [37]	72.0	63.1	77.9
PubMedCLIP [38]	60.1	80	72.1
BAN [39]	58.3	37.4	72.1
SAN [3]	54.3	31.3	69.5
Our model	**48**	**70.9**	**61.8**

From this table we can draw few insights. First, our model did very well as compared to the older papers that does not incorporate pre-training. Since due to resource and time constraints we could not pre-train our model on much larger medical datasets, it affected our model's generalization capability. There are some notable large vision language datasets that we can pre-train our model on in order to improve generalization and medical image and language feature extraction capabilities. Some notable large medical datasets are MedI-CaT [40], MIMIC-CXR [41], ROCO [42]. Second, our model did very well on close-ended questions, since we crafted our model mainly as a classifier and treated the Medical VQA task as a classification problem. Third, from the learning curve (figure 6) it seems that our model's accuracy was still increasing. So if we trained our model for few more epochs, we could see some more improvement in accuracy performance as well.

6. DISCUSSION & CONCLUSION

In this project, we have covered literature review on the current state of Medical Visual Question answering research. We have also proposed a novel architecture consisting of only transformer throughout every component of the visual question answering system. We have trained our model on the train set of VQA-RAD dataset and our model showed encouraging result on the test-set of VQA-RAD dataset. Furthermore, we believe pre-training on large medical vision-language dataset and additional tuning of model hyperparameters can greatly improve our model's performance and enhance its generalization and comprehension capabilities.

REFERENCES

1. Hasan, S. A., Ling, Y., Farri, O., Liu, J., Muller, H., & Lungren, M. (n.d.). *Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task*.

2. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., & Ge, Z. (2023). Medical Visual Question Answering: A Survey. *Artificial Intelligence in Medicine*, *143*, 102611. https://doi.org/10.1016/j.artmed.2023.102611

3. Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, *5*(1), Article 1. https://doi.org/10.1038/sdata.2018.251

4. Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., & Muller, H. (n.d.). *VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019*.

5. Kovaleva, O., Shivade, C., Kashyap, S., Kanjaria, K., Wu, J., Ballah, D., Coy, A., Karargyris, A., Guo, Y., Beymer, D. B., Rumshisky, A., & Mukherjee, V. M. (2020). Towards Visual Dialog for Radiology. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing* (pp. 60–69). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2020.bionlp-1.6

6. He, X., Zhang, Y., Mou, L., Xing, E., & Xie, P. (2020). *PathVQA: 30000+ Questions for Medical Visual Question Answering* (arXiv:2003.10286). arXiv. https://doi.org/10.48550/arXiv.2003.10286
7. Abacha, A. B., Datla, V. V., Hasan, S. A., & Muller, H. (n.d.). *Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain*.
8. Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., & Wu, X.-M. (2021). *SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering* (arXiv:2102.09542). arXiv. https://doi.org/10.48550/arXiv.2102.09542
9. Abacha, A. B., Sarrouti, M., Demner-Fushman, D., Hasan, S. A., & Müller, H. (n.d.). *Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain*.
10. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2016). *VQA: Visual Question Answering* (arXiv:1505.00468). arXiv. https://doi.org/10.48550/arXiv.1505.00468
11. Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556
12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*

(arXiv:2010.11929; Version 2). arXiv. http://arxiv.org/abs/2010.11929

14. 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* (arXiv:2103.14030). arXiv. https://doi.org/10.48550/arXiv.2103.14030

15. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

16. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* (arXiv:1406.1078). arXiv. https://doi.org/10.48550/arXiv.1406.1078

17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

18. Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 457–468). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1044

19. Yu, Z., Yu, J., Xiang, C., Fan, J., & Tao, D. (2018). Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(12), 5947–5959. https://doi.org/10.1109/TNNLS.2018.2817340

20. Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked Attention Networks for

Image Question Answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21–29. https://doi.org/10.1109/CVPR.2016.10

21. Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical Question-Image Co-Attention for Visual Question Answering. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper/2016/hash/9dcb88e0137649590b755372b040afad-Abstract.html
22. Kim, J.-H., Jun, J., & Zhang, B.-T. (2018). Bilinear Attention Networks. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html
23. Al-Sadi, A., Talafha, B., Al-Ayyoub, M., Jararweh, Y., & Costen, F. (n.d.). *JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain*.
24. Ren, F., & Zhou, Y. (2020). CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering. *IEEE Access*, *8*, 50626–50636. https://doi.org/10.1109/ACCESS.2020.2980024
25. Bansal, M., Gadgil, T., Shah, R., & Verma, P. (n.d.). *Medical Visual Question Answering at Image CLEF 2019- VQA Med*.
26. Zhou, Y., Kang, X., & Ren, F. (n.d.). *TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning*.
27. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.

https://doi.org/10.3115/1073083.1073135

28. Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., Zheng, Y., Liu, Y., Salim, F., Verspoor, K., Liang, X., & Chang, X. (n.d.). *FFA-IR: Towards an Explainable and Reliable Medical Report Generation Benchmark* (1.0.0) [dataset]. PhysioNet. https://doi.org/10.13026/CCBH-Z832

29. *Flaviagiammarino/vqa-rad · Datasets at Hugging Face*. (n.d.). Retrieved December 5, 2023, from https://huggingface.co/datasets/flaviagiammarino/vqa-rad

30. Liu, Y., Wang, Z., Xu, D., & Zhou, L. (2023). *Q2ATransformer: Improving Medical VQA via an Answer Querying Decoder* (arXiv:2304.01611). arXiv. http://arxiv.org/abs/2304.01611

31. Xu, L., Liu, B., Khan, A. H., Fan, L., & Wu, X.-M. (2023). *Multi-modal Pre-training for Medical Vision-language Understanding and Generation: An Empirical Study with A New Benchmark* (arXiv:2306.06494; Version 2). arXiv. http://arxiv.org/abs/2306.06494

32. Schuster, M., & Nakajima, K. (2012). *Japanese and Korean voice search*. 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079

33. *Probability for Machine Learning—Discover How To Harness Uncertainty With Python [v1.9&nbsp;ed.]*. (n.d.). Dokumen.Pub. Retrieved December 5, 2023, from https://dokumen.pub/probability-for-machine-learning-discover-how-to-harness-uncertainty-with-python-v19nbsped.html

34. Bridle, J. (1989). Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters. *Advances in Neural Information Processing Systems*, *2*. https://proceedings.neurips.cc/paper/1989/hash/0336dcbab05b9d5ad24f4333c7658a0e-Abstract.html

35. Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, *20*(3), 121–136. https://doi.org/10.1007/BF00342633

36. Hendrycks, D., & Gimpel, K. (2023). *Gaussian Error Linear Units (GELUs)* (arXiv:1606.08415). arXiv. https://doi.org/10.48550/arXiv.1606.08415

37. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U. D., & Jawahar, C. V. (2021). *MMBERT: Multimodal BERT Pretraining for Improved Medical VQA* (arXiv:2104.01394). arXiv. https://doi.org/10.48550/arXiv.2104.01394

38. Eslami, S., de Melo, G., & Meinel, C. (2021). *Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?* (arXiv:2112.13906). arXiv. https://doi.org/10.48550/arXiv.2112.13906

39. Nguyen, B. D., Do, T.-T., Nguyen, B. X., Do, T., Tjiputra, E., & Tran, Q. D. (2019). *Overcoming Data Limitation in Medical Visual Question Answering* (arXiv:1909.11867). arXiv. https://doi.org/10.48550/arXiv.1909.11867

40. Subramanian, S., Wang, L. L., Bogin, B., Mehta, S., van Zuylen, M., Parasa, S., Singh, S., Gardner, M., & Hajishirzi, H. (2020). MedICaT: A Dataset of Medical Images, Captions, and Textual References. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2112–2120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.191

41. *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports | Scientific Data*. (n.d.). Retrieved December 6, 2023, from https://www.nature.com/articles/s41597-019-0322-0

42. Pelka, O., Koitka, S., Rückert, J., Nensa, F., & Friedrich, C. M. (2018). Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In D. Stoyanov, Z. Taylor, S. Balocco,

R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, & P. Jannin (Eds.), *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (pp. 180–189). Springer International Publishing. https://doi.org/10.1007/978-3-030-01364-6_20