Enhancing Medical VQA with Self-Attention Based Multi-Model Approach

1rd Sakthivel V
Department of Computer Science
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
sakthiveloffcl@gmail.com

2st Bharathi Mohan G
Department of Computer Science
Amrita School of Computing
Amrita Vishwa Vidyapeetham
Chennai, India
g_bharathimohan@ch.amrita.edu

3th Elakkiya R
Department of Computer Science
BITS Pilani, Dubai Campus
Dubai, United Arab Emirates
elakkiya@dubai.bits-pilani.ac.in

2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10726054

***Abstract*—In the medical and healthcare fields, the integration of clinical images and question-answering systems is believed to be a powerful tool that has the capacity to change the pattern of diagnosing. The proposed model introduces an innovative approach using CNN and LSTM with self-attention for medical Visual Question Answering (VQA). Our approach serves the role of feature extraction and question embedding. The achieved results reveal the efficiency of the model in interpreting the VQA-RAD dataset, which mainly consists of radiology images. The baseline model obtained an accuracy of 70.71% without self-attention and 71.97% with self-attention. Additionally, the DenseNet Pretrained Model with self-attention scored an accuracy of 74.89%, and without self-attention, it achieved an accuracy of 73.22%, showing reliable performance against existing solutions. This paper contributes to the development of medical VQA systems and proposes avenues for future research and adjustments in the applications of healthcare systems.**

Keywords: LSTM, CNN, VQA-RAD, SELF-ATTENTION

I. Introduction

Medical imaging has enabled several advances in the field of healthcare, since it is used for diagnosis, surgery, and screening. By combining information from several sources, or multi-model data, doctors can reduce the number of incorrect diagnoses they make. Computer-aided diagnosis (CAD) has come a long way in the last few years because to the application of medical visual question answering (VQA), which blends computer vision (CV) with natural language processing (NLP).

Traditional medical image analysis includes the extraction of spatial data, including segmentation and detection. On the other hand, the need for advanced picture analysis has led to the development of complex paradigms like image captioning, which offers descriptive textual descriptions. Image captioning goes beyond simple item recognition by defining dynamic features, colors and textures, and connections between objects in the image.

VQA is an attempt to improve picture captioning into a more comprehensive form using recent advancements in computer vision. VQA improves comprehension of picture details by posing precise questions regarding the image content. While computer vision has a strong basis, VQA in medical imaging is still lacking, therefore there is room for development.

A typical VQA pipeline comprises four blocks: (i) a feature extractor for visual features from the image, (ii) a textual feature extractor from the question, (iii) an embedding module combining visual and textual modalities, and (iv) a prediction head.

The majority of modern medical VQA systems use deep learning methods, such as convolutional neural networks (CNNs) for visual feature extraction, recurrent neural networks (RNNs) for text embedding, and sophisticated processes like attention. Transformer network models, such as BERT, are commonly utilized for textual information representation in medical VQA due to the networks' effectiveness in natural language processing.

This work proposes a CNN-Bi-LSTM model with a self-attention architecture for VQA in medical pictures. During training, the model jointly analyzes both text and medical pictures, using LSTM for the text branch and CNN for the image branch. To generate the final response predictions, the extracted characteristics from both branches are merged. In the image branch, two CNN architectures are employed: one baseline CNN model and the other a DenseNet pretrained on CheXNet.

The remainder of the paper is organized as follows: Section II examines relevant research on VQA-RAD and VQA. The proposed VQA technique is presented in Section III. The VQA-RAD dataset undergoes thorough experimental investigation and comparisons in Section IV. Finally, Section V concludes by discussing possible future events and drawing conclusions.

II. Related Works

VQA sees a wide range of strategies, as detailed in [1][2], with the main focus being on modifying state-of-the-art general VQA models for medical use. The 2018 ImageCLEF-Med challenge offers a thorough overview of these techniques and their results [3]. Interestingly, writers in [4][5] often use complex annotation processes from general VQA, including SAN [6] and MCB [7], to create a coherent representation that links a picture to a question. It's interesting to note that these methods typically entail fine-tuning models that have already been trained on ImageNet, such as ResNet [8] or

VGG [9], in order to extract picture features that are specific to medical VQA. Unfortunately, the lack of readily available medical VQA datasets makes it useless to directly fine-tune such models using medical VQA data.

The study by Prasanna Kumar Rangarjan et al. [10] introduces a novel deep learning-based sentiment analysis framework for social media data. The authors significantly increased the accuracy of sentiment classification by utilizing deep neural networks and personalized sentiment dictionaries. The study's conclusions could offer helpful direction for improving medical question-answering systems' textual data analysis."

Recent developments in transformer-based abstractive summarization offer important new perspectives for improving the generation of text and understanding, as Bharathi Mohan Gurusamy et al. [11] investigated in their study. These observations are especially relevant to enhancing medical VQA systems, because it is essential to comprehend and summarize medical materials. Medical VQA systems can improve their understanding and response to intricate medical queries by integrating cutting-edge natural language processing (NLP) techniques such as transformer-based summarization.

Chantal Pellegrini et al. propose a novel way in [12] to address the current problem of having standardised radiology reporting, which will improve service delivery and accuracy in medical diagnosis. A new method, called Rad-ReStruct, uses a hierarchical annotation dataset, which is included as a benchmark. This dataset allows us to evaluate methods of automating structured reporting. According to their investigation, hi-VQA is the proposed hierarchical visual question answering technique which utilizes contextual information to write structured documentations. Their method shows competitive results on the underlying medical VQA benchmark VQARad. Researcher's highlight the insufficient existing research on automating structured reporting and the absence of public benchmark for evaluation. They characterize their work as a breakthrough in the process of structured radiology report population. Through scrupulous experimentation and confirmation of their model, they lay the foundation for future research in this discipline. Their results indicated the challenges and achievements in providing an automatic approach to medical imaging interpretation and generating standardized reports.

Bharathi Mohan et al. [13] used a stacked ensemble model that integrated Random Forest, Support Vector Machine, and Logistic Regression classifiers to show how well the model could identify between fake and real news articles. This methodological development emphasizes how using a variety of models in ensemble learning can improve categorization tasks. Similar to this, by combining predictions from several Visual Question Answering (VQA) models—including ones with self-attention mechanisms—a stacked ensemble technique may improve performance in the field of medicine.

A novel question-aware captioning model introduced in [14] to overcome the limitations of black-box language models (LMs) like GPT-3 in Vision Question Answering (VQA) tasks. Sometimes, the regular captions of images do not include details needed for VQA tasks. Therefore, we come up with the idea of question-aware captions to fill this gap. PROMPTCAP resolves this problem by creating individual captions from natural language prompts, so that LMs could comprehend and give appropriate answers. The authors suggest a pipeline that involves synthesis and filtration of samples with GPT-3's few-shot learning capability for the training of no-label PROMPTCAP. Experimental results show that PROMPTCAP, paired with GPT-3, achieve the state-of-the-art performance on the knowledge-based VQA tasks such as OK-VQA and A-OKVQA. Ablation studies have demonstrated that PROMPTCAP outperforms generic captions by a considerable margin proving its efficacy in improving LM-based VQA systems. Furthermore, PROMPTCAP's generalization ability is tested on WebQA which portrays its capability of working in different domains other than the main field. Overall, PROMPTCAP enables the bridge between visual information and LMs, resulting in precise and context-aware replies in VQA tasks at the end of the day.

The implementation of the Lottery Ticket Hypothesis (LTH) to trim the LXMERT model in [15], a two-stream vision-and-language (V+L) pretrained model, while aiming to optimize its performance in Visual Question Answering (VQA) tasks. The study was provoked by the motivation of resource-efficient models which investigate whether the smaller trainable subnetworks exist within LXMERT which are capable of being fine-tuned to the full performance. By applying magnitude pruning repeatedly, authors demonstrate that LXMERT can be pruned by 40%-60% in size while achieving only 3% loss in performance, which can make it possible to use LXMERT on resource-constrained devices for practical applications. The efficiency of the proposed pruning technique is confirmed via comparative analysis with DistillVLM. Moreover, the analysis is followed by a cost-benefit analysis showing a big accuracy loss starting from 50-60% pruning. These insights provide a rationale for the use of LTH in compressing V+L pretrained models to fulfill their purposes in the field.

ALign BEfore Fuse (ALBEF), a modern Vision-and-Language Pre-training (VLP) platform in [16], which overcomes the disadvantages of previous techniques, including the need for pre-trained object detectors and sensitivity to noisy data. In ALBEF, a detector-free approach is used, first encoder images and texts independently, and then fuse them through cross-modal attention. The ITC loss intermediate is contrastive that projects unimodal features from encoders and improves cross-modal learning as well as semantic understanding. Additionally, MoD is designed to engage larger uncurated datasets during training while not penalizing the model for offering output different from the correct one. Among various theoretical rationales, ALBEF also implies a maximization of mutual information between vision and language representations that are invariant to semantic-preserving changes. Experimental results on a variety of Vision-and-Language (V+L) tasks demonstrate ALBEF's performance superior to the state of the art, particularly in image-text retrieval, visual question answering, visual reasoning, visual entailment, and weakly-supervised visual grounding tasks. Under the circumstances, the ALBEF algorithm performs faster at inference and outperforms those models trained on bigger datasets, demonstrating its scalability and efficiency. Nevertheless, future research on data and model implications is required before the actual implementation be done because of the social impacts that could be a product of web data characteristics.

[17] On the dataset of Amazon reviews, BERT showed the highest accuracy as well as the lowest training and validation loss, making it the best model for sentiment analysis. Its exceptional performance was attributed to its capacity to extract semantic meaning and sophisticated contextual information from pre-training on a huge corpus.

The model [18] designed for tasks that combine vision and natural language, that is, challenges systems to understand the detailed semantics of images grounded in language. VisualBERT combines BERT with a pre-trained object proposals systems so that it could work with both image and text inputs together. VisualBERT's performance is tested on four vision-and-language tasks, such as visual question answering and visual commonsense reasoning, and reveals its superior or comparable ability in relation to the existing models. The model's key features, i.e., task-independent pre-training and early fusion of vision and language, are presented to be the core component for its success. Comprehensive analysis shows how VisualBERT implicitly correlates both sentences and regions from the input image by reaching high levels of grounding and capturing syntactic dependencies. Qualitative analysis demonstrates how VisualBERT is effective in fine-tuning alignments by layers of Transformers and resolving coreference. In short, VisualBERT presents an interesting method for joint representation of vision and language with further use in image-only tasks and large caption datasets in mind.

III. METHODOLOGY:

A. Dataset Collection:

The dataset utilized in this research is the VQA-RAD dataset, a specialized collection of radiology images designed explicitly for Visual Question Answering (VQA) tasks in the medical domain. The VQA-RAD dataset amalgamates visual and textual elements, featuring radiological images, associated textual questions, and corresponding answers (as depicted in Fig. 1). This dataset serves as a fundamental resource for training and evaluating our proposed multi-modal model for medical visual question answering.

B. Dataset Augmentation:

To enhance the diversity and robustness of the training dataset, we employ data augmentation techniques. The augment_data function is utilized, which randomly rotates images within a range of -20 to 20 degrees and applies random shifts to the width and height (up to 30 pixels). This augmentation strategy introduces variability into the dataset, aiding the model in learning invariant features.

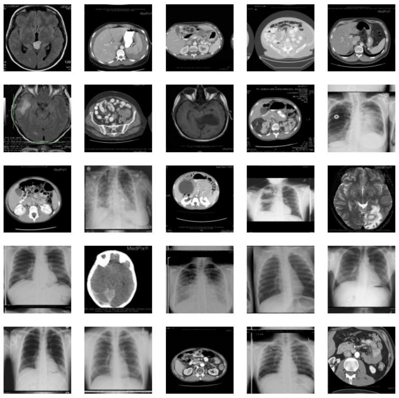


Fig. 1. Sample Dataset.

C. Dataset Preprocessing:

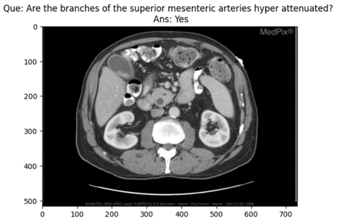
Image Processing: Images are read, decoded, and resized to a standardized dimension (e.g., 256x256 pixels) to ensure consistency. Pixel values are normalized to the range [0, 1] to facilitate convergence during model training Textual Sequence Handling: Textual sequences (questions) are tokenized and encoded to transform them into a format compatible with the multi-modal model architecture.

D. Train-Test Split:

The dataset is partitioned into training and validation sets:

Training Set: Used for optimizing model parameters and enhancing the model's ability to recognize patterns within the VQA-RAD data. Validation Set: Employed to assess the model's generalization and performance on previously unseen data.

E. Exploratory Data Analysis (EDA):



Que: Are the branches of the superior mesenteric arteries hyper attenuated?
Ans: Yes

Fig. 2. Question & Answer

1. Image Analysis: The dataset encompasses a variety of medical images, ranging from CT scans and MRI to X-rays, each with its unique size and modality (as illustrated in Fig. 1). Images display variability in their dimensions, with some reaching a maximum height and width of 1500 and 2321 pixels, respectively, while others have a minimum of 256 pixels for both dimensions. Despite these differences, all images maintain a consistent channel depth of 3, indicative of the RGB color format. To ensure uniformity and facilitate further analysis and modeling, it is advisable to resize the images to a standardized dimension, as shown in Fig. 3.

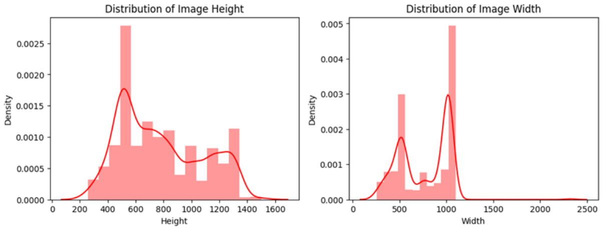


Fig. 3. Image Analysis

2. Question Analysis: Questions within the dataset exhibit a diverse range of lengths and complexities, reflecting the varied nature of inquiries posed by users. The character length of questions spans from a minimum of 12 to a maximum of 133, with an average of 37 characters per question. Similarly, the word count in questions varies, with an average of 6 words per question and some outliers extending beyond 10 words. Analysis based on the first one or two words of each question reveals common themes such as inquiries about imaging modality, specific organs, abnormalities, and attributes of the images, as depicted in Fig. 5.

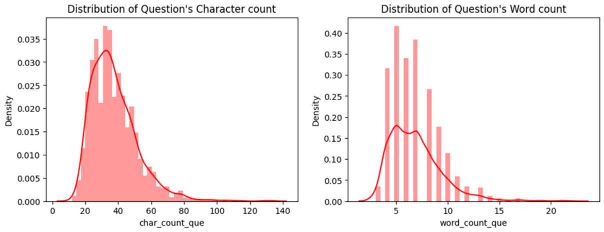


Fig. 4. Question Analysis

3. Answer Analysis: Answers provided in the dataset demonstrate a wide spectrum of lengths and types, reflecting the diversity of responses elicited by the questions, as shown in Fig. 5. The character length of answers ranges from 1 to 115, with an average of 8 characters per answer. Similarly, the word count varies, with an average of 1 word per answer and a maximum of 17 words. Notably, a significant proportion of answers are open-ended, allowing for diverse responses, although closed-ended responses are also prevalent. Common answers include medical terms related to organs, abnormalities, and imaging modalities, as well as affirmative or negative responses.

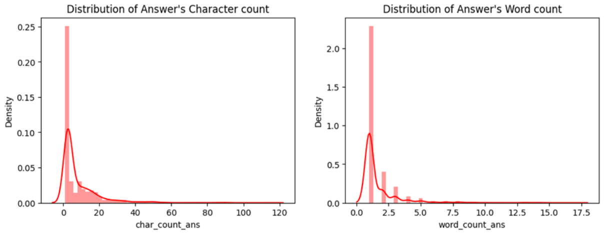


Fig. 5. Answer Analysis

4. Question & Answer Analysis: The analysis of questions and answers reveals several key patterns within the dataset. Firstly, a significant number of answers consist of ”yes” and ”no,” indicating prevalent binary response questions. Certain terms like ”brain,” ”CT,” ”fat,” ”axial,” ”left,” and ”right” are frequently repeated, suggesting common themes. Answers with 10 or more words are often associated with ”how” and ”what” questions, indicating a need for detailed explanations, as evident in Fig. 6. Questions starting with ”What” usually elicit answers related to attributes, while ”how” questions often result in numerical responses. ”Is” questions primarily generate ”yes” or ”no” answers, indicating straightforward inquiries.

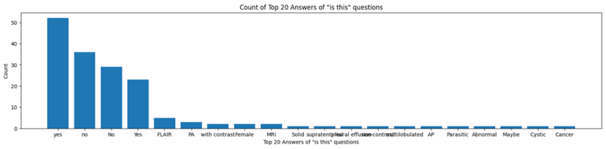


Fig. 6. Question & Answer Anaylsis

The analysis of questions and answers provides valuable insights into user queries and information retrieval needs within the medical imaging domain. By understanding the distribution of question types and corresponding answers, we can gain insights into common areas of interest among users. These insights inform the development of question-answering models tailored for medical imaging applications, enabling enhanced user interaction and more effective information retrieval.

IV. Architecture:

Our proposed model leverages a combination of Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and attention mechanisms for effective multi-modal learning. The architecture is designed to simultaneously process medical images and associated textual sequences.

A. *Image Branch*

The Image Branch was tested with two different CNN architectures. One CNN model serves as a baseline, where the image branch starts with a convolutional layer with 64 filters, followed by max-pooling and dropout for regularization. Two additional convolutional layers with 32 filters each are incorporated, interleaved with max-pooling and dropout operations. The resulting feature maps are then flattened for integration with the textual branch.

Another CNN model utilizes the pre-trained DenseNet121. The Image Branch begins by leveraging pre-trained weights from CheXNet, which are fine-tuned to our specific dataset. CheXNet CNN excels at capturing intricate patterns and abnormalities from medical images, providing robust visual representations. Following CheXNet, additional convolutional layers are employed for hierarchical feature extraction.

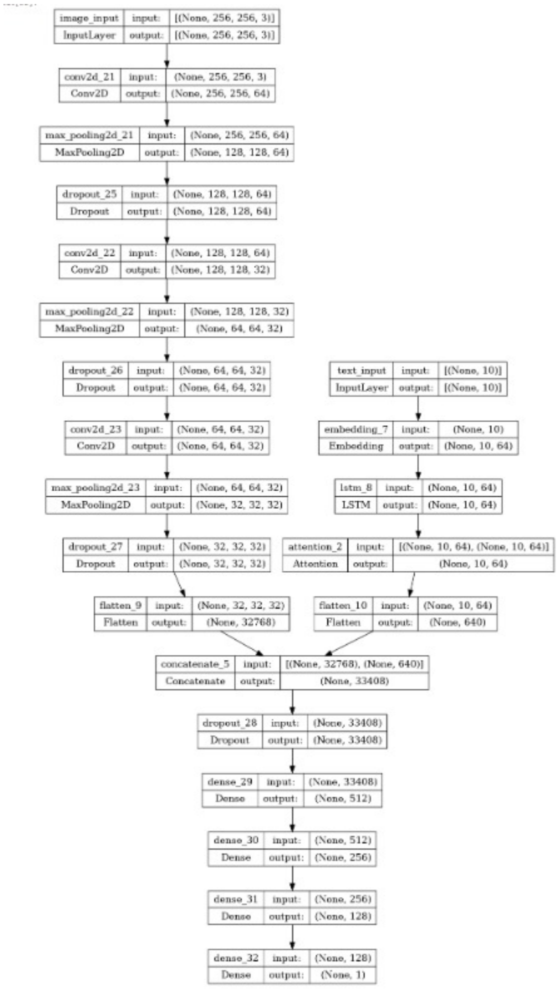

image_input | input: | [(None, 256, 256, 3)]
InputLayer | output: | [(None, 256, 256, 3)]
conv2d_21 | input: | (None, 256, 256, 3)
Conv2D | output: | (None, 256, 256, 64)
max_pooling2d_21 | input: | (None, 256, 256, 64)
MaxPooling2D | output: | (None, 128, 128, 64)
dropout_25 | input: | (None, 128, 128, 64)
Dropout | output: | (None, 128, 128, 64)
conv2d_22 | input: | (None, 128, 128, 64)
Conv2D | output: | (None, 128, 128, 32)
max_pooling2d_22 | input: | (None, 128, 128, 32)
MaxPooling2D | output: | (None, 64, 64, 32)
dropout_26 | input: | (None, 64, 64, 32)
Dropout | output: | (None, 64, 64, 32)
text_input | input: | [(None, 10)]
InputLayer | output: | [(None, 10)]
conv2d_23 | input: | (None, 64, 64, 32)
Conv2D | output: | (None, 64, 64, 32)
embedding_7 | input: | (None, 10)
Embedding | output: | (None, 10, 64)
max_pooling2d_23 | input: | (None, 64, 64, 32)
MaxPooling2D | output: | (None, 32, 32, 32)
lstm_8 | input: | (None, 10, 64)
LSTM | output: | (None, 10, 64)
dropout_27 | input: | (None, 32, 32, 32)
Dropout | output: | (None, 32, 32, 32)
attention_2 | input: | [(None, 10, 64), (None, 10, 64)]
Attention | output: | (None, 10, 64)
flatten_9 | input: | (None, 32, 32, 32)
Flatten | output: | (None, 32768)
flatten_10 | input: | (None, 10, 64)
Flatten | output: | (None, 640)
concatenate_5 | input: | [(None, 32768), (None, 640)]
Concatenate | output: | (None, 33408)
dropout_28 | input: | (None, 33408)
Dropout | output: | (None, 33408)
dense_29 | input: | (None, 33408)
Dense | output: | (None, 512)
dense_30 | input: | (None, 512)
Dense | output: | (None, 256)
dense_31 | input: | (None, 256)
Dense | output: | (None, 128)
dense_32 | input: | (None, 128)
Dense | output: | (None, 1)

Fig. 7. CNN - Architecture

B. Textual Branch

The textual branch commences with an embedding layer to convert tokenized sequences into dense vectors. A Bidirectional LSTM layer with 64 units is employed to capture temporal dependencies in the textual data. Furthermore, a Multi-Head Self-Attention mechanism is integrated to enhance the model's ability to focus on relevant parts of the textual input.

These networks consist of input, forget, and output gates (i_t, f_t, o_t), a candidate cell state ($\tilde{C}t$), and a cell state (C_t). The input gate determines relevant input ($i_t = \sigma(Wxix_t + W_{hi}h_{t-1} + b_i)$), while the forget gate controls previous cell state retention ($f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$). The output gate regulates output based on the current cell state ($o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$). The candidate cell state is generated from current input and previous hidden state ($\tilde{C}t = \tanh(Wxcx_t + W_{hc}h_{t-1} + b_c)$), and the cell state is updated considering the gates and candidate cell state ($C_t = f_t \times C_{t-1} + i_t \times \tilde{C_t}$). Finally, the hidden state is produced as the output, influenced by the cell state and output gate ($h_t = o_t \times \tanh(C_t)$).

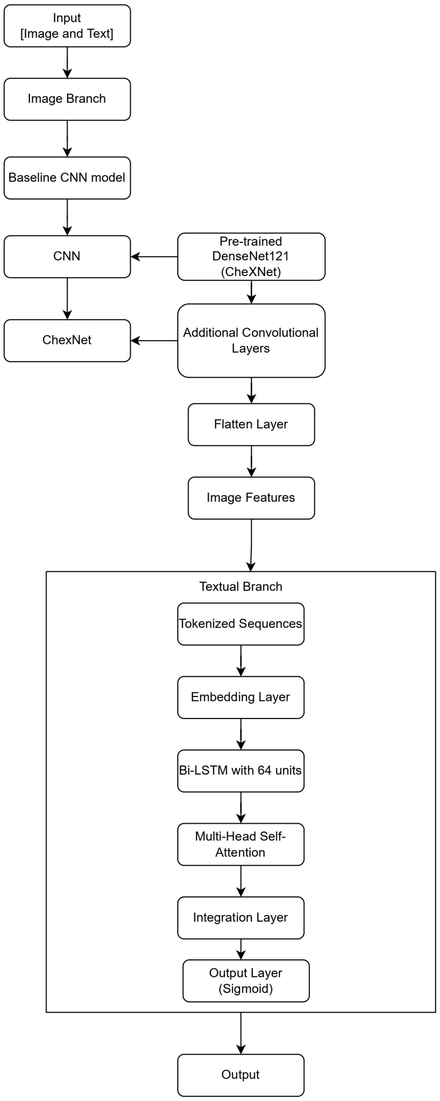


Fig. 8. Overall Architecture

C. Fusion and Dense Layers

The outputs from the image and textual branches are concatenated and passed through dropout layers to mitigate overfitting. Subsequent dense layers with rectified linear unit (ReLU) activation functions help the model learn complex representations. The final output layer, comprising a single neuron with a sigmoid activation function, produces the binary classification output.

D. Model Compilation and Training

The model is compiled using the Adam optimizer and binary cross-entropy loss function, suitable for binary classification tasks. Training is performed on the augmented dataset with batch sizes of 32, utilizing a buffer size of 500 for shuffling. The training data is further pre-fetched for optimized processing.

E. *Validation*

The model's performance is assessed on a separate validation dataset. Evaluation metrics such as accuracy are monitored to gauge the model's ability to generalize to unseen data.

V. Results

In this research, we investigated the effectiveness of integrating self-attention mechanisms into convolutional and recurrent neural network architectures for multimodal learning tasks. We compared the performance of baseline models against models enhanced with self-attention mechanisms, both on their own and in conjunction with a pre-trained DenseNet model.

A. *Baseline Model*

Our baseline model consisted of convolutional and recurrent neural network components for processing image and text inputs, respectively. The model achieved a validation accuracy of 70.71% after 15 epochs of training.

Introducing self-attention mechanisms into the baseline model resulted in an improvement in performance, with a validation accuracy of 71.97% achieved after 15 epochs, as shown in Fig. 9.

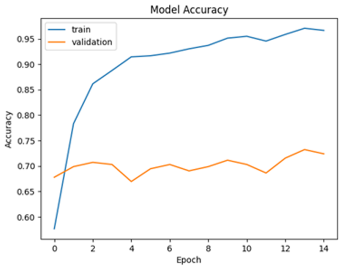


Fig. 9. LSTM + BaseLine Model

B. *DenseNet Pretrained Model*

We also explored the integration of a pre-trained DenseNet model, originally trained on the CheXNet dataset, into our architecture. Despite the powerful feature extraction capabilities of DenseNet, the model yielded a validation accuracy of 73.22%, which did not surpass the baseline model's performance.

Similar to the baseline model, incorporating self-attention mechanisms into the DenseNet pretrained model led to improved performance, achieving a validation accuracy of 74.89% after 15 epochs. (refer to Fig. 11)

Overall, our experiments demonstrate that the integration of self-attention mechanisms can enhance the performance of multimodal learning architectures, even when combined with powerful pre-trained models like DenseNet. While there were improvements observed, further research is warranted to explore the optimal integration and tuning of self-attention mechanisms in various multimodal architectures for different tasks and datasets.

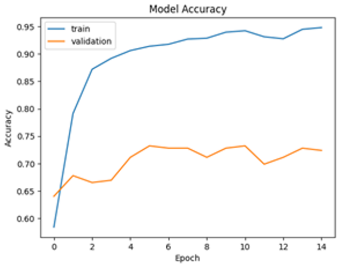
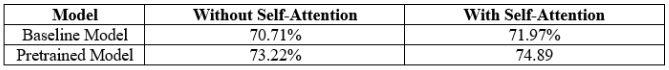


Fig. 10. LSTM + Densenet Modedl



Model	Without Self-Attention	With Self-Attention
Baseline Model	70.71%	71.97%
Pretrained Model	73.22%	74.89

Fig. 11. Validation Accuracy Comparison

VI. Conclusion

In this study, we explored multiple models for VQA. Our CNN-LSTM Fusion Model achieved the highest accuracy at 74.895%, showcasing the effectiveness of combining spatial and sequential features. The Basline model without self attention gained accuracy of 70.71% , Baseline Model with Self-Attention achived accuracy of 71.97%.Next set of models are DenseNet Pretrained Model with and without self-Attention and gained accuracy of 73.22% and 74.89% respectively These findings highlight the potential of multi-modal approaches for VQA in medical imaging. Future research can focus on refining attention mechanisms and exploring interpretability for clinical applicability.

References

[1] Abacha, A. B., Gayen, S., Lau, J. J., Rajaraman, S., & Demner-Fushman, D. (2018, September). NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain. In CLEF (working notes) (pp. 1-10).

[2] 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), 1-10.

[3] Hasan, S. A., Ling, Y., Farri, O., Liu, J., Müller, H., & Lungren, M. P. (2018, September). Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. In CLEF (Working Notes).

[4] Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21 (pp. 52-59). Springer Berlin Heidelberg.

[5] Zhou, Y., Kang, X., & Ren, F. (2018, September). Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering. In CLEF (working notes) (pp. 1-11).

[6] Yang, Zichao, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. "Stacked attention networks for image question answering." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 21-29. 2016.

[7] Fukui, Akira, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[9] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[10] P. K. Rangarjan, B. M. Gurusamy, G. Muthurasu, R. Mohan, G. Pallavi, S. Vijayakumar, and A. Altalbe, ”The social media sentiment analysis framework: deep learning for sentiment analysis on social media,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 3, pp. 3394-3405, Jun. 2024. DOI: 10.11591/ijece.v14i3.pp3394-3405

[11] B. M. Gurusamy, P. K. Rengarajan, and P. Srinivasan, ”A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer,” International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no. 6, pp. 6663–6672, Dec. 2023.

[12] Pellegrini, C., Keicher, M., Özsoy, E., & Navab, N. (2023, October). Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 409-419). Cham: Springer Nature Switzerland.

[13] B. M. G, H. R, J. K, S. V, S. V. P and V. MS, ”Fake News Detection Using a Stacked Ensemble of Machine Learning Models,” 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2024, pp. 1165-1169, doi: 10.1109/IDCIoT59759.2024.10467326.

[14] Hu, Yushi, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. ”Promptcap: Prompt-guided task-aware image captioning.” arXiv preprint arXiv:2211.09699 (2022).

[15] Hashemi, Maryam, Ghazaleh Mahmoudi, Sara Kodeiri, Hadi Sheikhi, and Sauleh Eetemadi. ”LXMERT Model Compression for Visual Question Answering.” arXiv preprint arXiv:2310.15325 (2023).

[16] Li, Junnan, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. ”Align before fuse: Vision and language representation learning with momentum distillation.” Advances in neural information processing systems 34 (2021): 9694-9705.

[17] Rangarajan, Prasanna Kumar & Gurusamy, Bharathi Mohan & R, Elakkiya & M, Charan & M, Rithani. (2023). Automated Sentiment Classification of Amazon Product Reviews using LSTM and Bidirectional LSTM. 1-6. 10.1109/EASCT59475.2023.10393514.

[18] Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. ”Visualbert: A simple and performant baseline for vision and language.” arXiv preprint arXiv:1908.03557 (2019).