Deep Learning Based Approaches for Medical Visual Question Answering using Transformer and Mixup

Anurag Jaiswal
2021csm1002@iitrpr.ac.in
Dr. Deepti R. Bathula
bathula@iitrpr.ac.in

Department of Computer Science & Engineering
Indian Institute of Technology Ropar
Punjab, India

Abstract

In Artificial Intelligence, Medical VQA has become a very active area of research due to the need for AI systems that can answer medical images such as X-rays, MRI scans, and ultrasounds. This AI system can help medical practitioners overcome the burden of examining multiple images and help patients better understand their medical condition. Visual Question Answering is an extension of image captioning tasks in today's world which lies in the area of computer vision and NLP. Visual question answering on natural images has become so popular, which enables the interest in medical domain applications. As medical images differ from natural images and due to the scarcity of medical datasets, complex deep learning models are not very fruitful; researchers apply several techniques, such as the Attention mechanism-based model, which attends to several parts of images that are important concerning questions. Due to the availability of fewer medical datasets, research also used transfer learning to leverage the usage of pre-trained models for better generalization. This thesis project will focus on implementing a deep learning-based model that can classify the answers from the possible set of answers to such questions related to medical images, which can help doctors overcome their burden and allow patients to understand their medical condition better and benefit society.

1 Introduction

Insufficient medical practitioners worldwide have increased the need for software models that correctly answer medical images such as MRI scans, x-rays, CT scans, ultrasounds, etc. Medical VQA can be seen as an extension of natural image question answering. Med VQA recently became an active research area due to an increase in the need for such a system that can help in making such medical decisions that can change human life and benefit society. It relies mainly on natural language processing (NLP), deep learning, and computer vision techniques. Building an automatic VQA system can help doctors overcome the burden of examining multiple medical images. Compared to natural images, medical images are challenging to understand, and due to the scarcity of available medical datasets, the problem becomes more challenging. Building AI systems with fewer data that can answer questions related to medical images such as X-rays, MRI, and ultrasounds becomes interesting. This thesis project will focus on implementing a model that can classify the answers from the possible answers to such questions related to medical images. VQA is a kind of multimodal-based framework that involves the areas of vision and NLP.VQA task consists of a task of multi-class image classification. VGG16, ResNet, and DenseNet are baseline architectures used to extract image features, and LSTM (Long short-term Memory)or pre-trained BERT model is used to extract question features. Using ResNet and DenseNet, the pre-trained BERT model leverages the transfer learning concept, providing better generalization as the baseline models were not very accurate on natural image datasets.

Therefore researchers focus on the attention mechanism, which basically meant to pay attention to the relevant artifacts of the image or the question. As humans process data or simple text by doing different parts of it similarly in this mechanism paying attention to certain parts of input text will improve performance in case of long input sequences. RNNs such as LSTM help us to improve performance for temporal data. For spatial data after performing the convolution operation instead of using the fully connected layer, the output of convolution layers can be considered as a specific path of volume. As we know, we can trace back to the particular patch of the original image passed through CNN. So the output feature map after convolution, considering one particular path part of that depth volume, can provide spatial information. Using this spatial information, we can assign different weights to different parts of the image and create a context vector that helps to understand which part of the image is more relevant with respect to the question. Attention can be of several types as Hard and Soft attention. In hard attention, choosing a particular part of the image is the only focus for generating a specific output that is stochastic and non-differentiable, as soft attention will not choose a specific part of the image rather than assign weights to every aspect of the image which helps to generate more deterministic and differentiable output. The other categorization is Local and Global Attention. All the input parts are chosen for attention rather than local attention to global attention. The inputs are chosen only from a neighborhood area or the region of interest for attention. There are several other kinds of attention, such as self-attention. Such an attention mechanism has become very popular in image captioning tasks for natural images.

2 Literature Survey

Several models are used to address the Visual Question Answering Problem. One of the baseline models combines LSTM and Image features that we get out of a Convolution Neural Network. Given an image to a CNN such as VGG net and received a fully connected layers representation at the end of Convolution neural network(CNN). Pass each word through different timesteps of an LSTM by concatenating these two features and sending them to fully connected layers to make final predictions. This baseline model is discussed in the paper [1]—Visual Question Answering. Hence CNNs and RNNs Attention is used to mix matches to solve such multi-modal problems. Another baseline model developed in 2015 [2] for visual question answering used the concept of a Bag of words for creating the frequency histogram and used as text input image features extracted out of one of the layers of the Convolution neural network. Several efforts going in the direction were in CVPR of 2016 [3] by authors published a paper known as the Stacked Attention Network for visual question answering, which is very similar to attention models that worked on top of feature representation which is obtained after applying convolution. Hence these attention models help to take particular convolution layers to feature maps, then map those spatial locations in the original image and get the feature vectors corresponding to each part of the image. Then based on the attention, another level of engagement is again performed on the image feature vectors of different parts. The first attention layer estimates which part of the image to focus on and which part it focuses on wherever there are usual concepts in the question. After using the attention on the images, it leads to the scenario that we can also apply attention to questions. Hence another work was proposed in 2016 NeuroIPS [4] in which the author named the model as Hierarchical Co-Attention model. In this model, co-attention signifies that it attends to parts of the image and questions, it gives attention to both question and the image, and the hierarchy suggests embedding at a word level of the question, then adding a phrase level of the question, and then sentence level, which is termed as a question. So extract the word-level embeddings,phrase-level embeddings, and question-level embeddings. In most of the text-related problems, to a large extent, we have a vocabulary of words. Each word in question corresponds to 1 hot vector on vocabulary, which is input for a particular comment. After extracting the word, phrase, and question level embedding, apply co-attention to the question and corresponding image at each level. So Attention maps help us understand the model we are focusing on. To get the phrase-level embedding, take embeddings of each caption word, and perform convolution operation across those inputs with multiple filters of different widths. In that direction, the work done by Microsoft Research is visual question-answering and Image Captioning tasks using Bottom-Up and Top-Down Attention. However, in this approach, all the image features are divided into uniform grids, which are then weighted by top-down attention.

Recent research in the field of Visual question answering follows the observation to determine the kind of phenomena leading to self-supervised or semi-supervised learning. One of the efforts in this direction is to overcome the prior language problem mentioned by authors [5]. They focused on the inherent data bias problem that the dataset has. Because some answers are present in high frequency, the model is biased towards those sets of solutions. Their simple observation for solving the problem is that they answer only relevant questions. The simple and intuitive statement is a question that can only be answered based on its relevancy. The authors proposed a QICE model named Question Image Correlation Estimation to estimate the relevance between the question and the image. They generated an equal number of relevant and irrelevant question image pairs from the dataset, signified them as the probabilistic term, and added them to the loss function, which acts as a regularizer. The task is to train a question-image correlation estimation model to predict the ground-truth label of each question-image pair by minimizing cross-entropy loss. Hence it reduces the loss and the accuracy in the case of relevant question image pairs and increases the loss and decreases the accuracy in the case of irrelevant question image pairs. This way, the authors ensure the model performs well on test data. They proposed a unified self-supervised vqa framework where the loss function contains two terms: the first relates to the cross-entropy of a loss, and the second is to regularization, which estimates the question dependency. Researchers also used transfer learning [6] due to the scarcity of medical datasets available. They used the pre-trained Bert model, which is trained on thousands of radiology scans and shows improvement in the medical visual question-answering task. Researchers also used different data augmentation techniques, such as the Mixup [7] approach, which facilitates data and helps keep the model simple while improving accuracy. They applied the mixup data augmentation technique by mixing two images with the mixing parameter chosen from a beta distribution with hyperparameter alpha. As mixing questions does not make sense, the model generates the predictions for both questions whose corresponding images are mixed. Then combine the predictions using the mixing parameter λ chosen from a beta distribution to generate the final prediction. The researchers also propose a novel method for visual question answering (VQA) that combines multi-modal factorized bilinear pooling and co-attention learning. The proposed method combines multi-modal factorized bilinear pooling, which learns the interaction between visual and textual features in a joint embedding space, with co-attention learning, which focuses on essential regions of the image and words in the question simultaneously. This approach allows for more effective modeling of the complex interactions between the two modalities. The authors note that VQA in surgical scenes is particularly challenging due to the complex and dynamic nature of the operating room environment. The proposed approach uses a transformer model that leverages both visual and textual features to generate answers to questions about surgical procedures.

3 Proprosed Problem Statement

In the medical domain, given a natural language question related to medical Images, the task is to provide an accurate answer from a possible set of solutions.

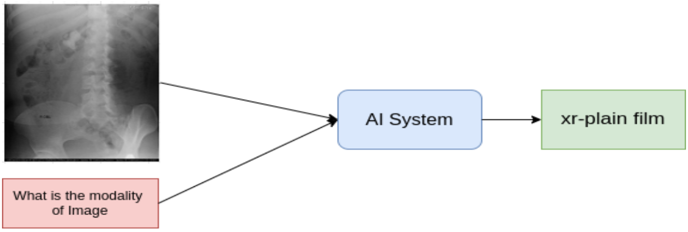


Figure 1: Med-VQA Challenge

Such an AI system's performance is crucial in the medical domain as medical practitioners do not want a misclassification by the model as human life depends on it. Hence, implementing such AI models becomes crucial and subjective for their usage in a real-world scenario. The increase in interest in the medical artificial intelligence domain for medical care and health showcases the need to improve the performance of existing VQA models is a challenging task.

4 Dataset Description

ImageCLEF is an organization that provides the dataset for which contains MRI, X-Ray, CT-Scan, and Ultrasound. The motivation is to support multilingual users, visual media, and cross-language annotation. The other available datasets are VQA-RAD and Sysu.

ImageCLEF2019 Dataset			
Description	Train	Validation	Test
Unique Images	3200	500	500
Unique Questions	247	186	138
Unique Answers	1552	470	470

As the training set has many unique answers and many unique answers appear in the dataset only once, we limit the number of unique classes to only those which appear in the dataset more than a particular threshold (which we set to 5). This way, the dataset will reduce to 178 unique classes, which is more balanced.

5 Proposed Methodologies

As briefly discussed the attention mechanism in the introduction and Literature section, we were inspired by the MFB (Multi-modal Factorized Bilinear) Pooling technique with co-attention for vqa task on natural images. We also used a multilayer transformer encoder and BERT to perform visual question-answering tasks. We experiment with multiple Convolutional Neural Networks in model architecture to extract the images' features. Such CNNs are VGG19, Resnet-50, and VGG19 with Global Average Pooling. We also experimented with multiple Recurrent Neural Networks to extract text features in model architecture. Such RNNs are BERT and Bio-Clinical BERT.

5.1 Extraction of Image Features for Visual Question Answering Task

Extract Features from the pre-trained VGG19 model from each layer.

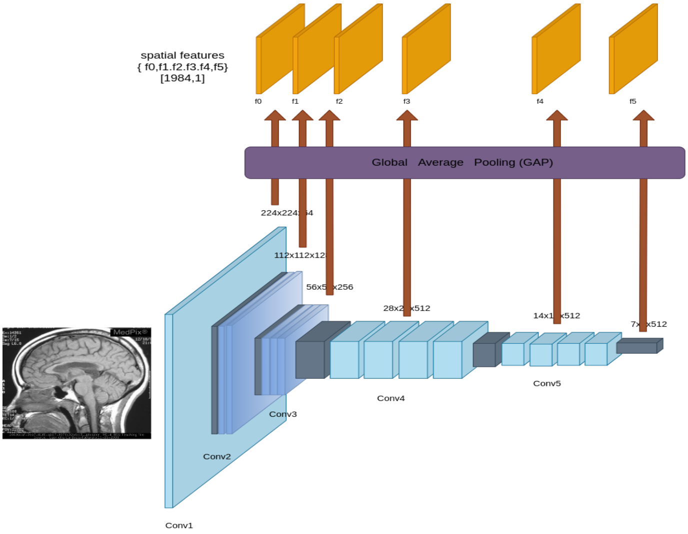


Figure 2: VGG19+GAP(Global Average Pooling)

Perform global average pooling (GAP) that allows to concat the features which help in providing more spatial information as compared to VGG19 or Resnet50 only.

5.2 MFB with Co-attention architecture for Med-VQA

MFB technique combines the feature vectors of different modalities, such as visual features and text features.MFB can be viewed as a two-step process; in the first step, the features from various modalities are expanded to a multi-faceted or large dimension-based field, and in the second step use, element-wise multiplication is for integration.

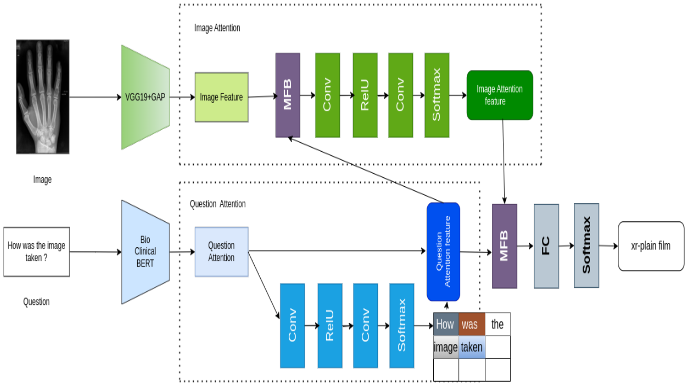


Figure 3: MFB with Co-attention base architecture for Med-VQA

Hence we experimented with multiple CNNs and RNNs for MFB with Co-attention architecture for Medical VQA Task.

5.3 visualBERT Transformer based architecture for Med-VQA

The VisualBERT is a transformer that uses BERT and object proposal models for processing language and images together. It works by analyzing the words in a sentence using BERT and the visual features of objects in the image using object proposal models. It then combines this information in several layers to create a comprehensive representation that grasps both the language and the objects. The VisualBERT applies in many areas that demand an understanding of language and images.

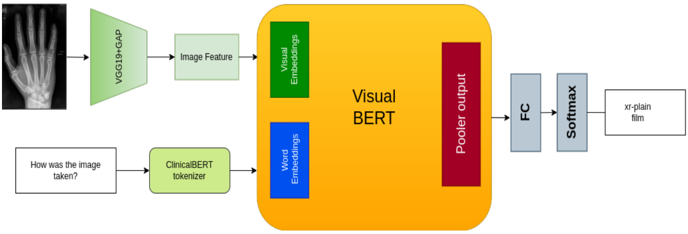


Figure 4: VisualBERT base architecture for Med-VQA

5.4 VQA Mixup

This technique is a way of augmenting data that adapts the MixUp approach for use in visual question-answering (VQA) tasks. The process involves generating a new image, Vmix, by taking a weighted combination of original images, V_x and V_y, and their corresponding ground truth labels, A_x and A_y, using the equations:

$$V_{mix} = \lambda V_x + (1 - \lambda) V_y$$
$$A_{mix} = \lambda A_x + (1 - \lambda) A_y$$

The value of the mixing parameter λ, which ranges from 0 to 1, is randomly picked from a Beta distribution with α as the hyperparameter. This approach is intended to improve the performance of the VQA model by creating new training samples with slight variations, leading to better generalization and robustness.

5.4.1 Metrics for Vqa-Mixup

Given a new image V and corresponding questions Q_x and Q_y, a VQA model (M) generates predictions P_x and P_y for Q_x and Q_y, respectively. The number of classes is large. We evaluate top1 and top5 accuracy. We also mixed the predictions based on mixing parameter (λ) for top1,top5 accuracy, and bleu score.

$$\text{Predictions: } P_x = M(V_x, A_x), P_y = M(V_y, A_y)$$

$$\text{MixupLoss: } Loss_{mix} = \lambda \mathcal{L}(P_x, A_x) + (1-\lambda)\mathcal{L}(P_y, A_y)$$

$$\text{Top1 Accuracy: } top1Acc_{mix} = \lambda Correct(P_x, A_x) + (1-\lambda)Correct(P_y, A_y)$$

$$\text{Top5 Accuracy: } top5Acc_{mix} = \lambda Correct(P_x, A_x) + (1-\lambda)Correct(P_y, A_y)$$

$$\text{Bleu Score: } bleu_{mix} = \lambda bleu(P_x, A_x) + (1-\lambda)bleu(P_y, A_y)$$

5.5 CrossEntropy Loss & Contrastive Loss based Combined Loss

During a model's training, the primary goal is to minimize a loss function that measures the difference between the original and predicted output. When dealing with multi-class classification problems, CrossEntropy Loss is a popular loss function. However, in certain s, such as visual question answering, the model must learn a combined embedding area for both the image and text inputs. In such cases, Contrastive Loss can be used in addition to CrossEntropy Loss as an additional loss function. Therefore, we conducted various experiments using combined loss based on contrastive and cross-entropy loss to improve the model's performance.

6 Experimental Result

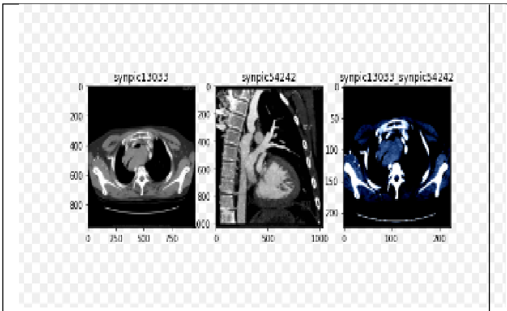
To improve the model's performance, we adopted various techniques to enhance the training process. We saved the features of original and mixed images in secondary memory, which helped achieve smoother and faster model training. This approach also facilitated better utilization of available resources and minimized the computational load.

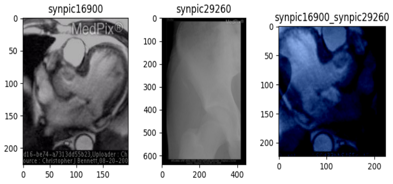
The model's training was carried out using the original, mixed, and combination of both datasets. This was done to evaluate the effect of mixing images on the model's performance. We also visualized the mixed images to understand the mixed dataset better. These experiments were performed by considering the mixing of images as a data augmentation technique.

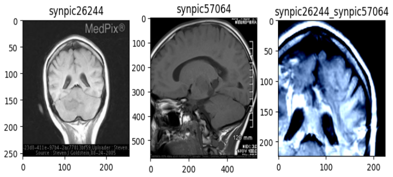
We implemented the transformer base architecture, which uses its self-attention module to create a dependency between visual and text-based tokens. We passed the pooler output of the visualBERT encoder to a fully connected layer for classification. We also incorporate contrastive loss in addition to cross-entropy loss by minimizing the distance between a particular image and text embedding and maximizing the distance between that image and other text embeddings in one batch. These loss functions helped improve the test accuracy of the model on the test dataset.

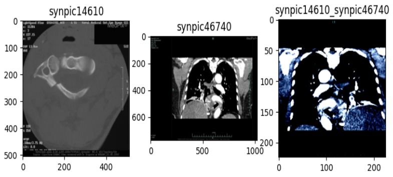
Overall, the results of these experiments demonstrate that incorporating mixup data augmentation and using transformer-based architectures along with different loss functions can significantly enhance the performance of medical visual question-answering systems. This can positively impact society by improving the accuracy of medical image analysis and aiding in better patient care.

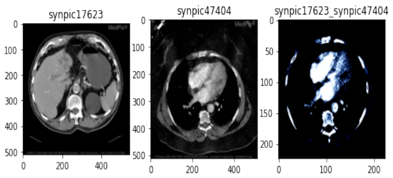
6.1 Visualization of Mixed Images

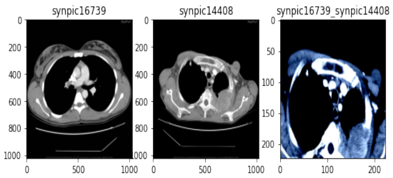






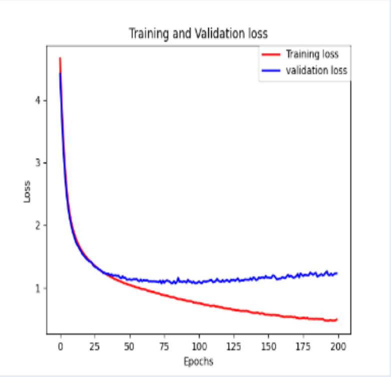
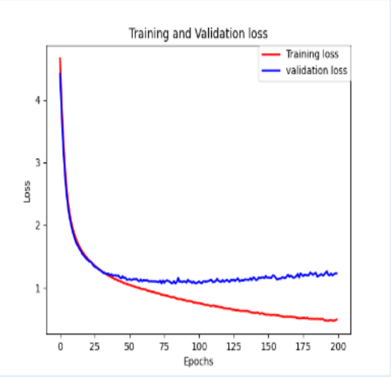
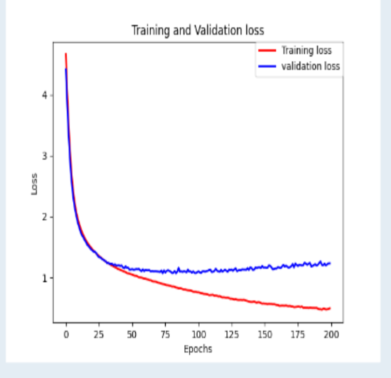
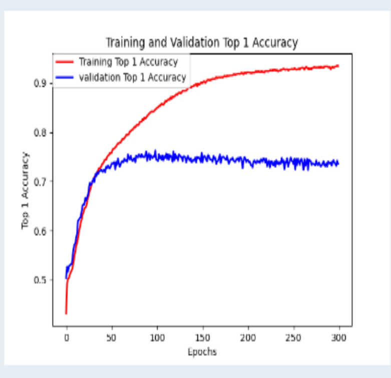
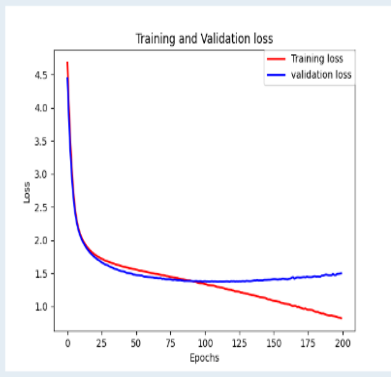
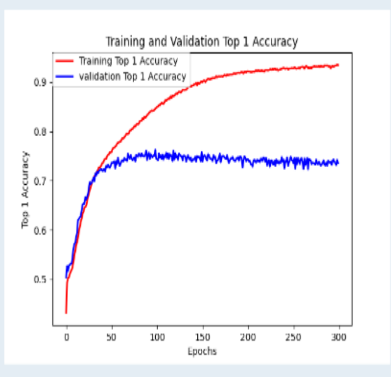


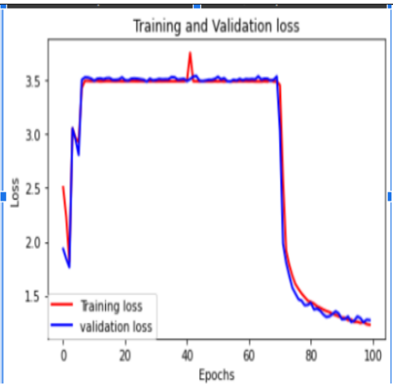




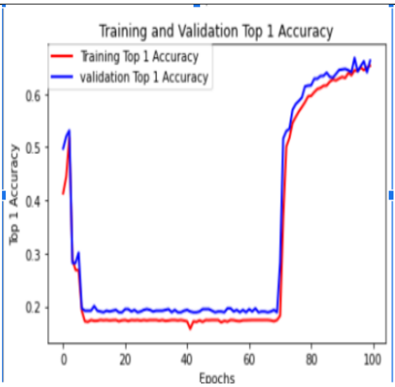
Bleu Score (Training)	Best Validation Top1 Accuracy	Best Validation Top5 Accuracy	Bleu Score (Validation)	Testing Accuracy	Modality	Plane	Organ	Abnormality
0.5101	0.7565	0.9535	0.41	0.54	0.784	0.65	0.68	0.048
0.52	0.76	0.9676	0.40	0.55	0.784	0.65	0.68	0.048
0.46	0.73	0.9245	0.39	0.536	0.784	0.65	0.68	0.048
0.4655	0.668	0.9055	0.40	0.425	0.784	0.65	0.68	0.048
0.4655	0.72	0.91	0.41	0.52	0.784	0.65	0.68	0.048
0.3548	0.6875	0.9085	0.41	0.46	0.784	0.65	0.68	0.048
0.35	0.7565	0.9535	0.41	0.46	0.56	0.624	0.52	0.048
0.3828	0.7565	0.9535	0.41	0.50	0.784	0.65	0.68	0.048
0.4476	0.7565	0.9535	0.3918	0.536	0.728	0.672	0.64	0.088
0.5101	0.6860	0.9410	0.41	0.55	0.784	0.65	0.68	0.088

Architecture	Classes Threshold	Original	Mixed	Original + Mixed	Mixup Coefficient	Learning Rate	Loss	Training Top1 Accuracy	Training Top5 Accuracy
VGG19+GAP + BERT + MFB + Co-attention	5	1	0	0	0	1.00E-04	CrossEntropy	0.8979	0.9576
VGG19+GAP + ClinicalBERT + MFB + Co-attention	5	1	0	0	0	1.00E-04	CrossEntropy	0.90	0.96
VGG19+GAP + ClinicalBERT + MFB + Co-attention	5	1	1	1	0.1	1.00E-04	CrossEntropy	0.81	0.89
VGG19+GAP + BERT + MFB + Co-attention	5	0	1	0	0.2	1.00E-04	CrossEntropy	0.8325	0.9403
VGG19+GAP + BERT + MFB + Co-attention	5	1	1	1	0.2	1.00E-04	CrossEntropy	0.80	0.87
VGG19+GAP + BERT + MFB + Co-attention	5	0	1	0	0.4	1.00E-04	CrossEntropy	0.6444	0.8251
VGG19+GAP + visualBERT	5	1	0	0	0	1.00E-04	CrossEntropy	0.65	0.89
VGG19+GAP + visualBERT	5	1	1	1	0.2	1.00E-04	CrossEntropy	0.68	0.88
VGG19+GAP + visualBERT	5	1	0	0	0	1.00E-04	CrossEntropy +Contrastive	0.76	0.95
VGG19+GAP + visualBERT	5	1	1	1	0.2	1.00E-04	CrossEntropy + Contrastive	0.65	0.89

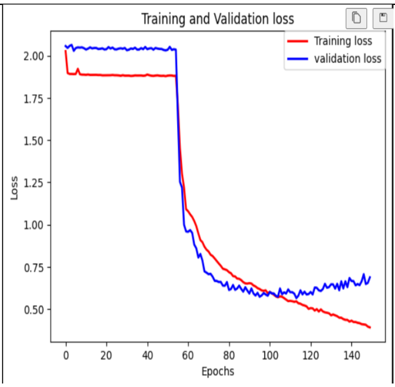
Training/Validation Loss	**Training/Validation Accuracy**
 VGG19+GAP+BERT+MFB	 VGG19+GAP+BERT+MFB
 VGG19+GAP+ClinicalBERT+MFB	 VGG19+GAP+ClinicalBERT+MFB
 VGG19+GAP+BERT+MFB+ Original+Mixed	 VGG19+GAP+BERT+MFB+ Original+Mixed



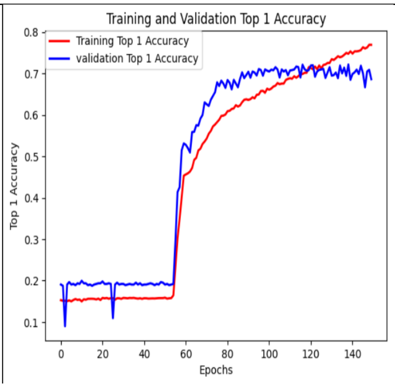
VGG19+GAP+visualBERT



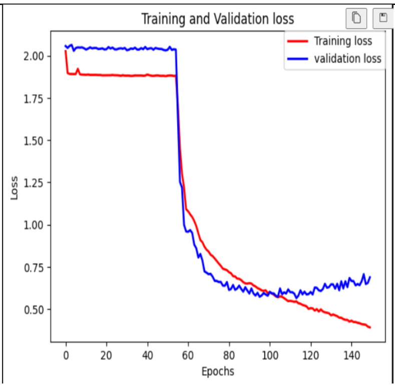
VGG19+GAP+visualBERT



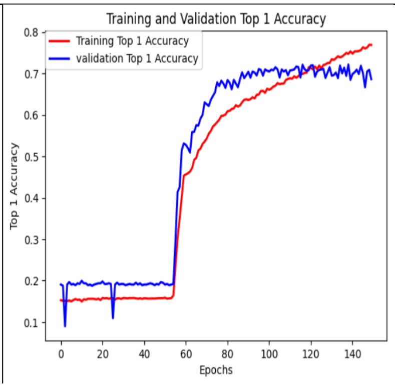
Combined Loss on VGG19+GAP+visualBERT



Combined Acc on VGG19+GAP+visualBERT+



Combined Loss on VGG19+GAP+visualBERT+ Original+Mixed



Combined Acc on VGG19+GAP+visualBERT+ Orginal+Mixed

7 Conclusion

In conclusion, this thesis aimed to implement such architecture, which improves the model's performance. Through various experiments, we implemented the visualBERT encoder, which takes the visual embedding from VGG19+GAP-based pre-trained architecture and text embedding from the pre-trained Bert model and includes the implementation of different loss functions such as CrossEntropy Loss and Contrastive Loss, as well as the use of image pairing data augmentation-based techniques like mixup and attention mechanism, we were able to enhance the model's accuracy as compared to MFB with Co-attention architecture for medical VQA. We also improved the MFB with co-attention-based architecture by using the bio-clinical Bert as compared to the Bert model, which is most commonly used on natural images.

The use of MVQA has the potential to benefit society by helping doctors assess medical images and allowing patients to understand their medical conditions better. The improvements made to the MVQA model through this thesis can contribute to the advancement of medical image analysis and help improve patient care.

8 Future Work

Medical Visual Question Answering (MVQA) is an active field of research that has the potential to revolutionize medical diagnosis and treatment by allowing physicians and healthcare professionals to more efficiently and accurately access medical information. We can use domain knowledge to improve the model performance further. Future work should focus on creating larger, more diverse datasets that can capture a broader range of medical knowledge and visual information.

References

[1] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

[2] Zhou, Bolei, et al. "Simple baseline for visual question answering." arXiv preprint arXiv:1512.02167 (2015).

[3] Yang, Zichao, et al. "Stacked attention networks for image question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." Advances in neural information processing systems 29 (2016).

[5] Zhu, Xi, et al. "Overcoming language priors with self-supervised learning for visual question answering." arXiv preprint arXiv:2012.11528 (2020).

[6] Khare, Yash, et al. "MMBERT: multimodal BERT pretraining for improved medical VQA." 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021.

[7] Singh, Jitender, Dwarikanath Mahapatra, and Deepti R. Bathula. "MEDICAL VQA: MIXUP HELPS KEEP IT SIMPLE." (2022).

[8] Seenivasan, L., Islam, M., Krishna, A.K. and Ren, H., 2022, September. Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII (pp. 33-43). Cham: Springer Nature Switzerland.