Visual Question Answering for Medical Images with Explainable AI

Deepananth K 195001027
Jayakrishnan S V 195001040
BE CSE, Semester 7

Dr. S Kavitha
Supervisor

Project Review: 1 (10 February 2023)
Department of Computer Science and Engineering
SSN College of Engineering

Abstract

Visual Question Answering(VQA) combines the fields of Natural Language Processing and Computer Vision to generate answers for the questions about the given input image. The project uses images from ImageCLEF 2019 VQA-Med Dataset and these images are complex to analyze and are of low resolution. VQA involves fusion of features extracted from both image and corresponding question, and the fused feature vector is then used for training a Neural Network based model. The trained model is then used for generating answers for the given input image and question. Explainable AI(XAI) is a recently trending domain which analyzes the Machine Learning and Deep Learning models for the given input and gives us the result of the analysis which is the desired explanation. Combining XAI with VQA for the medical images gives analysis supports the generated answer with justifications.

1 Introduction

Artificial Intelligence has grown exponentially over the past 10-15 years. The intelligent models or agents have solved many real-world problems and were able to learn or identify patterns among different kinds of data and provide the desired output. Now moving on to the next phase of learning, where the model tries to answer the questions asked by the user related to some data provided along with the question. Visual Question Answering(VQA) is one such emerging task in the field of Artificial

Intelligence and Computer Vision that aims to generate answers for the given questions by looking into the given image which corresponds to the question. VQA can be applied to various types of images like Natural Images, Medical Images or Cartoon Images. In this project, we aim to use different types of medical images like radiology images, CT scans, MRI scans etc., along with relevant questions and try to generate answers. Features are extracted from both image and question. The features are fused and are used to train a Neural Network architecture. The trained Neural Network architecture is used to generated answer for the input image and question. In order to justify the answer, some explanations are needed. There should be some features that correspond to the answer that is generated. Such features can be analyzed and identified using XAI tools.

1.1 Motivation

The medical domain is one, where there are new viruses and new diseases and also the one that needs faster analysis of different patients' conditions. Applying Artificial Intelligence techniques in the Medical field is more effective, where deep analysis of problems van be performed with the help of different AI techniques or algorithms. Visual Questions Answering in the medical domain would help doctors to analyze and get in-depth knowledge of medical images. Also, the doctors can submit their queries and get the required information [8]. Not only doctors, but even patients could use these VQA tools to get answers to their questions. Instead of searching and reading unknown articles from various websites, they can use these tools to get some required information. There are Visual Question Answering models[2, 3, 4, 5] for the medical domain that could generate answers for questions, but they do not give any justification for the predicted outcome. To overcome this limitation, the proposed model uses a Explainable AI (XAI) technique to justify the outcome of the VQA model.

1.2 Problem statement

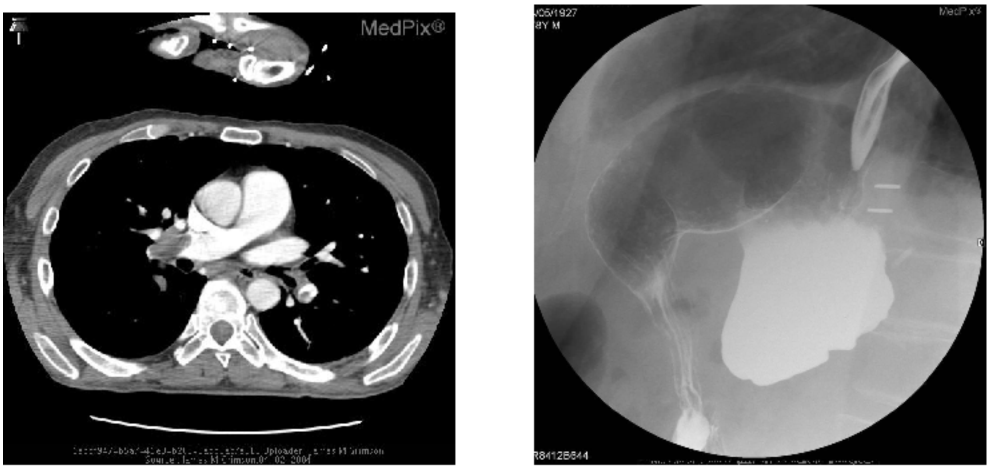
The aim of this project is to build an efficient VQA model that generates answers to questions related to Medical Images using deep learning techniques. In addition, the reason behind the generated answer has to be analyzed using Explainable AI tools like LIME, SHAP to provide explanations on the outcome.

1.3 Input

The input is the medical images of different modalities, disease types, planes etc., and their relevant questions.

1.4 Output

For the given image and the query the proposed system predicts the answer which will be validated using Explainable AI technique. A sample image, query with the corresponding answer is shown in Figure 1.



(g) **Q**: which organ system is shown in the ct scan? **A**: lung, mediastinum, pleura

(h) **Q**: what is abnormal in the gastrointestinal image? **A**: gastric volvulus (organoaxial)

Figure 1: Sample Images and Questions with Corresponding Answers from ImageCLEF 2019 VQA-Med Dataset (Image IDs: synpic191614, synpic28495)

1.5 Objectives

- To collect and analyze the dataset of Visual Question Answering from CLEF Forum.
- To compare and validate different transformer models for text feature extraction.
- To build an efficient VQA model that generates the answers to the questions related to the given Medical Image.
- To analyze the generated answer, Explainable AI tools have to be used for more explanations.

2 Literature survey

This section discusses about various research papers on Visual Question Answering with its techniques and limitations and about XAI techniques used for Deep Learning algorithms. The discussions are summarized into Table 1.

Table 1: Literature Survey

Paper Title	Methodology	Limitations
MoBVQA: A Modality based Medical Image Visual Question Answering System [1]	A CNN is trained for the different modalities like X-Ray, MRI, CT, Ultrasound **Dataset**: ImageCLEF 2019 VQA-Med **Answer Generation**: CNN Classifier **Analysis**: Accuracy-60.8, BLEU Score-63.4	Only modality based questions are considered
An Encoder-Decoder model for visual question answering in the medical domain [2]	**Dataset**: ImageCLEF 2019 VQA-Med **Image Feature Extraction**: DenseNet-121 **Question Feature Extraction**: LSTM **Feature Fusion**: Feature Concatenation **Answer Generation**: Fully Connected Neural Network **Analysis**: Accuracy-60.8, BLEU Score-63.4	For each query, entire image needs to be looked up, while a attention based mechanisms could be used to look up only the question centric regions

Visual question answering in the medical domain based on deep learning approaches: A comprehensive study [3]	The Questions are classified into 4 categories and multiple models are trained for each type of question **Dataset**: ImageCLEF 2019 VQA-Med **Image Feature Extraction**: VGGNet16 **Answer Generation**: Ensemble of Classification models **Analysis**: Accuracy-60.8, BLEU Score-63.4	All models built for each question categories are classification models which is completely a black-box approach
MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain [4]	**Dataset**: ImageCLEF 2019 VQA-Med, PathVQA **Image Feature Extraction**: ResNet152 **Question Feature Extraction**: BERT **Feature Fusion**: Multimodal Compact Bilinear Pooling (MCB) **Answer Generation**: LSTM **Analysis**: Accuracy-63.6	Answer Generation is based on LSTM and comparatively transformer based models like BERT work efficiently
Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering [5]	**Dataset**: ImageCLEF 2019 VQA-Med, PathVQA **Image Feature Extraction**: Inception-ResNet-V2 **Question Feature Extraction**: Bi-LSTM **Feature Fusion**: Concatenation **Answer Generation**: Fully Connected Layer **Analysis**: Accuracy-63.6	Answer Generation is based on LSTM and comparatively transformer based models like BERT work efficiently

Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain [6]	**Dataset:** Red Lesion Endoscopy data **XAI tools:** LIME, SHAP, CIU A CNN is trained using the dataset. XAI tools are then used for visualization in terms of heatmap. The result of visualization is then compared	
LISA : Enhance the explainability of medical images unifying current XAI techniques [7]	**Dataset:** COVID-19 Dataset **XAI tools:** LIME, SHAP, Anchors **Other XAI techniques:** Integrated Gradients Transfer Learning is used for the detection of COVID-19. The XAI tools LIME, SHAP Anchor and Integrated Gradient techniques' results are combined to give explanations.	

Inference

- From the analysis, it is inferred that no explanations were provided by the existing VQA models.
- For this project, the ImageClef 2019 VQA-Med dataset is chosen since it has a diverse categories of questions.
- There are various models for extracting features from the image. In this project, VGGNet and a custom CNN are to be used.
- To extract question features, transformer based models like BERT is to be used.
- For answer generation, BERT is to be used.
- For XAI, LIME and SHAP are to be used.

3 Proposed system

The proposed system aims to develop a VQA model for the ImageCLEF 2019 VQA-Med Dataset using VGGNet, BERT, LIME, SHAP.

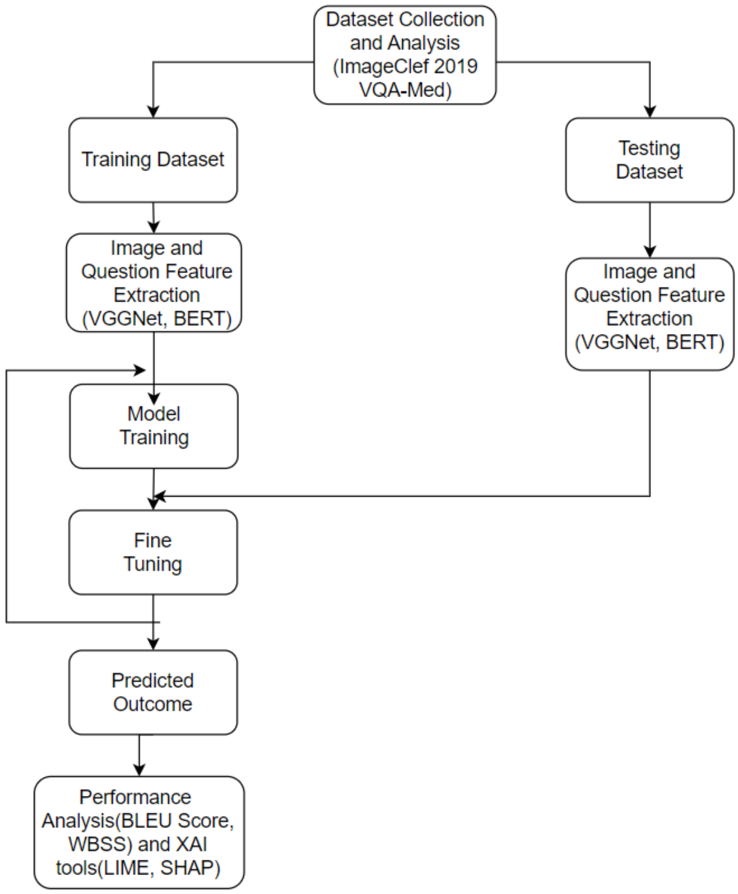


Figure 2: System Design

The system design is split into four modules. They are:

1) Dataset Collection and Analysis
2) Feature Extraction
3) VQA Model Building
4) Performance Analysis and XAI

3.1 Dataset Collection and Analysis

ImageClef 2019 VQA-Med Dataset.

3.2 Feature Extraction

A custom CNN and pre-trained VGGNet are used to extract features from the image. The text features are to be extracted using transformer models like BERT.

3.3 VQA Model Building

This module involves developing a VQA model that takes the fused feature vector as input and gives us the corresponding answer for the fused features. Neural Network based models such as LSTM, BERT and/or any similar models could be used.

3.4 Performance Analysis and XAI

The performance of the VQA model can be analyzed in terms of BLEU Score or WBSS. XAI tools such as SHAP and LIME are to be used for justifying the generated answers using heatmaps.

4 Feasibility study

4.1 Availability of Dataset

ImageClef 2019 VQA-Med Dataset
Dataset Link: https://www.aicrowd.com/clef_tasks/29/task_dataset_files?challenge_id=220

4.2 Timeline

Review	Module	Jan	Feb	Mar	Apr
Review 1	Data Collection and Analysis				
	Image Feature Extraction				
Review 2	Text Feature Extraction				
	VQA Model Building				
Review 3	Performance Analysis (BLEU Score, WBSS) and XAI (LIME, SHAP)				

4.3 Hardware and Software Requirements

High processing computers with GPUs are required for training the model faster and efficiently. Machine Learning and Deep Learning libraries like Tensorflow, Pandas, Numpy, Pytorch, CV2 are needed.

5 Implementation & Results

For review 1, the dataset collection & analysis and image feature extraction are completed. The flow of the work carried out for review 1 is depicted in Figure 3.

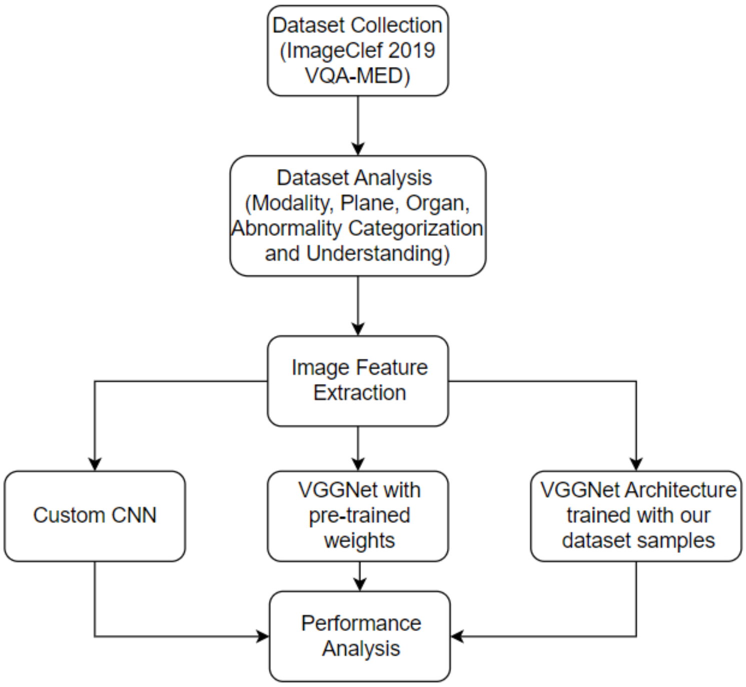


Figure 3: Image Feature Extraction

5.1 Dataset Collection & Analysis

Many Datasets are available for the desired tasks and they contain many Medical Images along with many relevant questions for each image. A ImageCLEF task has been posted for VQA for Medical Images and the dataset(VQA-Med 2019) is under AI Crowd. The dataset contains different medical images and their corresponding Question-Answer pairs. There are 3200 training medical images. The questions are categorized into four major types - Modality, Plane, Organ System and Abnormality. There are totally 12,792 Question-Answer pairs.

Dataset	Question Category	No. of Questions	No. of Classes
Training Dataset	Modality	3200	44
	Plane	3200	15
	Organ System	3200	10
	Abnormality	3192	1484
Testing Dataset	All	500	-

Table 2: Dataset Analysis

A text file for each category of questions is given in the dataset which includes the Question-Answer pair along with the image ID which corresponds to the name of the image file.
The validation set contains 500 images and 2000 Question-Answer pairs. The test set consists of 500 images and 500 Question-Answer pairs.The result of the analysis is given in the Table 2.

A sample image along with four types of question and corresponding answers is shown in Figure 4.

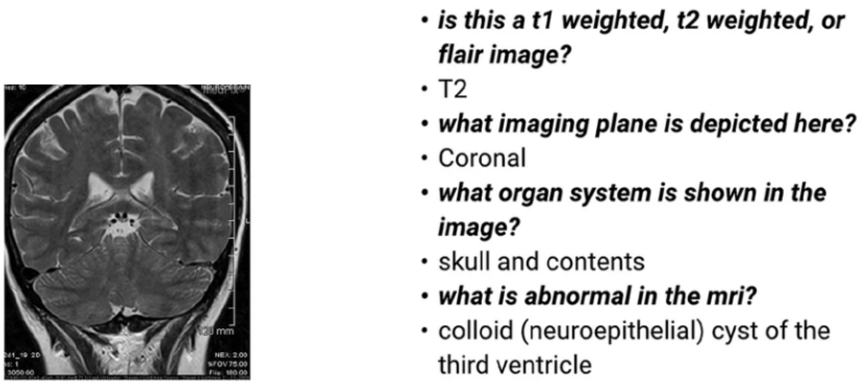


Figure 4: A Sample Image and QA pair from Dataset(Image ID: synpic16994)

5.2 Image Feature Extraction using CNN and VGGNet

Over 3200 samples that belongs to organ category are taken from the ImageClef 2019 VQA-Med training dataset for training and 500 samples are taken for testing.

The summary of the Custom CNN and VGGNet models are shown in Figure 5 & 6.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 222, 222, 32)	896
conv2d_5 (Conv2D)	(None, 220, 220, 32)	9248
max_pooling2d_2 (MaxPooling 2D)	(None, 110, 110, 32)	0
conv2d_6 (Conv2D)	(None, 108, 108, 32)	9248
conv2d_7 (Conv2D)	(None, 106, 106, 32)	9248
max_pooling2d_3 (MaxPooling 2D)	(None, 53, 53, 32)	0
flatten_1 (Flatten)	(None, 89888)	0
dense_3 (Dense)	(None, 400)	35955600
dense_4 (Dense)	(None, 300)	120300
dense_5 (Dense)	(None, 10)	3010

Total params: 36,107,550
Trainable params: 36,107,550
Non-trainable params: 0

Figure 5: Model Summary of Custom CNN

Model: "vgg16"

Layer (type)	Output Shape	Param #
input_3 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000

Total params: 138,357,544
Trainable params: 0
Non-trainable params: 138,357,544

Figure 6: Model Summary of VGGNet

The comparison of performances of Custom CNN, Pre-trained VGGNet and VGGNet trained with ImageClef 2019 VQA-Med dataset are summarized in the Table 3.

DNN Techniques	Training Accuracy	Testing Accuracy
Custom CNN	0.9906	0.512
Pre-trained VGGNet	0.7125	0.681
VGGNet Trained with our Dataset	0.5706	0.586

Table 3: Comparison of various feature extraction techniques

The training and testing performance of Custom CNN, Pre-trained VGGNet and VGGNet trained with dataset are shown in Figure 7 - 9.

```
In [84]: feature_extraction_model.fit(train_images_cnn,class_labels_cnn,epochs=15,callbacks=[early])

         Epoch 1/15
         100/100 [==============================] - 42s 421ms/step - loss: 20.3516 - accuracy: 0.4141
         Epoch 2/15
         100/100 [==============================] - 44s 437ms/step - loss: 1.0038 - accuracy: 0.6956
         Epoch 3/15
         100/100 [==============================] - 51s 512ms/step - loss: 0.3508 - accuracy: 0.9022
         Epoch 4/15
         100/100 [==============================] - 64s 637ms/step - loss: 0.1476 - accuracy: 0.9638
         Epoch 5/15
         100/100 [==============================] - 64s 644ms/step - loss: 0.0813 - accuracy: 0.9837
         Epoch 6/15
         100/100 [==============================] - 75s 754ms/step - loss: 0.0409 - accuracy: 0.9906
         Epoch 6: early stopping

Out[84]: <keras.callbacks.History at 0x7fb5d6ecba60>

In [97]: feature_extraction_model.evaluate(test_images_cnn,class_labels_cnn_test)

         16/16 [==============================] - 6s 309ms/step - loss: 2.9579 - accuracy: 0.5120

Out[97]: [2.9579203128814697, 0.5120000243186951]
```

Figure 7: Custom CNN (Training, Testing)

```
In [95]: feature_extraction_model.fit(train_images_cnn,class_labels_cnn,epochs=15,callbacks=[early])

Epoch 1/15
100/100 [==============================] - 227s 2s/step - loss: 2.3566 - accuracy: 0.4897
Epoch 2/15
100/100 [==============================] - 230s 2s/step - loss: 1.7020 - accuracy: 0.5869
Epoch 3/15
100/100 [==============================] - 225s 2s/step - loss: 1.4965 - accuracy: 0.6269
Epoch 4/15
100/100 [==============================] - 217s 2s/step - loss: 1.2757 - accuracy: 0.6641
Epoch 5/15
100/100 [==============================] - 219s 2s/step - loss: 1.1086 - accuracy: 0.6784
Epoch 6/15
100/100 [==============================] - 219s 2s/step - loss: 1.0020 - accuracy: 0.7125
Epoch 6: early stopping

Out[95]: <keras.callbacks.History at 0x7f4626064040>

In [103]: feature_extraction_model.evaluate(test_images_cnn,class_labels_cnn_test)

16/16 [==============================] - 35s 2s/step - loss: 1.3849 - accuracy: 0.6820

Out[103]: [1.3849315643310547, 0.6819999814033508]
```

Figure 8: Pre-trained VGGNet (Training, Training)

```
In [73]: feature_extraction_model.fit(train_images_cnn,class_labels_cnn,epochs=15,callbacks=[early])

Epoch 1/15
100/100 [==============================] - 575s 6s/step - loss: 4.1537 - accuracy: 0.3731
Epoch 2/15
100/100 [==============================] - 543s 5s/step - loss: 1.6779 - accuracy: 0.4672
Epoch 3/15
100/100 [==============================] - 398s 4s/step - loss: 1.5833 - accuracy: 0.4772
Epoch 4/15
100/100 [==============================] - 379s 4s/step - loss: 1.4571 - accuracy: 0.5263
Epoch 5/15
100/100 [==============================] - 374s 4s/step - loss: 1.4026 - accuracy: 0.5447
Epoch 6/15
100/100 [==============================] - 372s 4s/step - loss: 1.3143 - accuracy: 0.5706
Epoch 6: early stopping

Out[73]: <keras.callbacks.History at 0x7fe683e17370>

In [82]: feature_extraction_model.evaluate(test_images_cnn,class_labels_cnn_test)

16/16 [==============================] - 12s 742ms/step - loss: 1.3404 - accuracy: 0.5860

Out[82]: [1.3404346704483032, 0.5860000252723694]
```

Figure 9: VGGNet weights Trained with our Dataset (Training, Testing)

5.3 Visualization

To develop an efficient feature extraction model and to choose the correct number of layers, visualization(using heatmap) techniques are used. Figure 10 - 12 shows the Heatmap of every techniques implemented such as Custom CNN, Pre-trained VGGNet and VGGNet with weights Trained with our Dataset.

```
In [120]: sample_no = 105
          print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
          _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

synpic47017 --> skull and contents
1/1 [==============================] - 0s 41ms/step
1/1 [==============================] - 0s 62ms/step
9.611272e-06
```

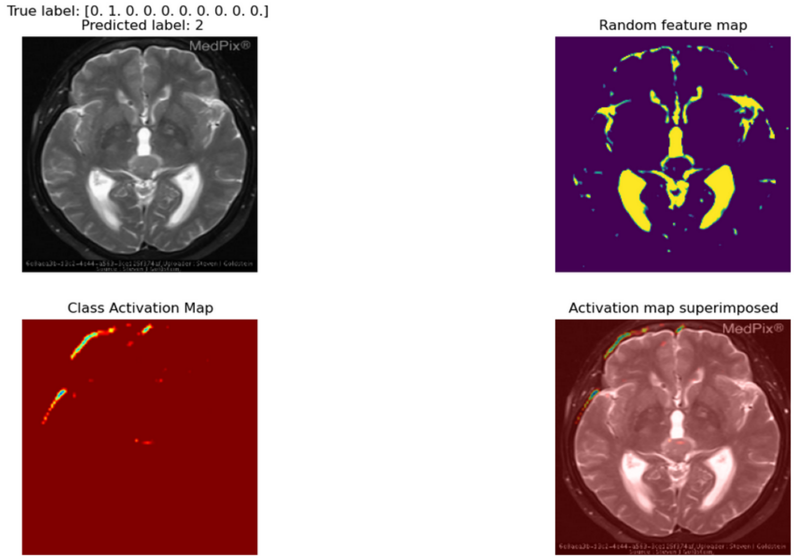


Figure 10: Custom CNN

```
In [124]: sample_no = 105
          print(OrganImageMapping_test[0][sample_no],"-->",OrganImageMapping_test[2][sample_no])
          _=show_random_sample(OrganImageMapping_test[0][sample_no],class_labels_cnn_test[sample_no])

synpic47017 --> skull and contents
1/1 [==============================] - 0s 77ms/step
1/1 [==============================] - 0s 116ms/step
1.4341319e-07
```

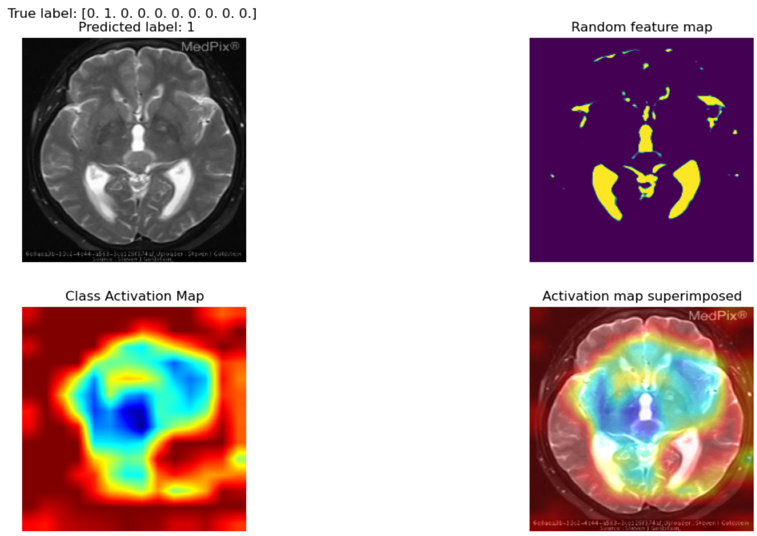


Figure 11: Pre-trained VGGNet

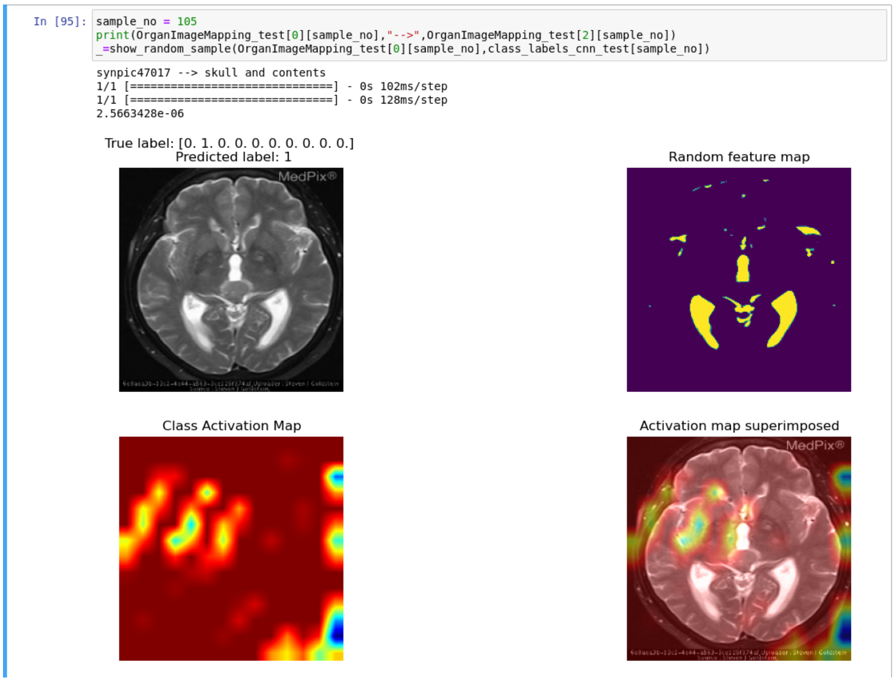


Figure 12: VGGNet weights Trained with our Dataset

6 Implementation Related Challenges

The challenges faced during the implementation are listed below:

1. **Dataset Complexity:** The dataset consists of images that are of different modalities(MRI, CT, X-Ray, etc.,) and different planes(AP, Lateral, Axial, etc.,). These affect the process of feature extraction using CNN and hence the resulting features are not as desired.

2. **Overfitting:** There exists a high imbalance in the dataset which may make the model overfit on some particular classes.

3. **The text present in images affects the feature extraction:** The Deep Learning Model CNN takes the text present on the image as an important feature. In general, most of the medical images may contain some text. This reduces the effectiveness of the extracted feature. A sample image containing text and its heatmap are shown in Figure 13.

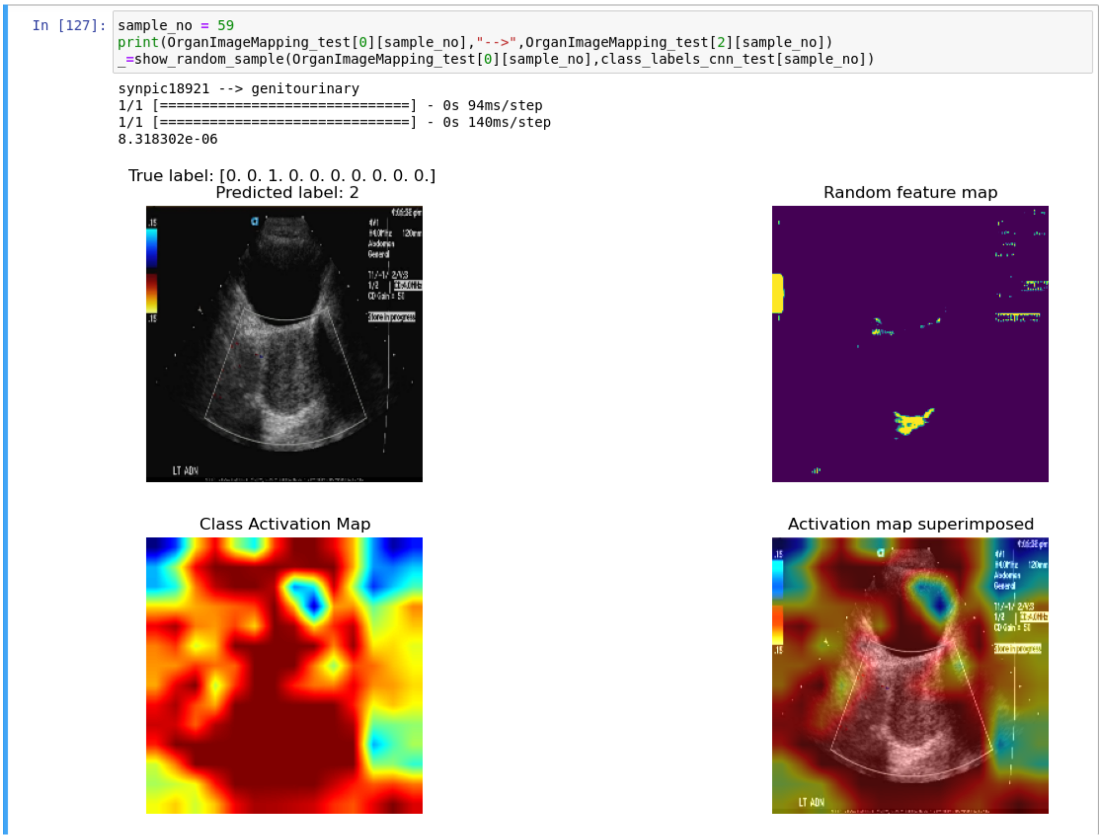


Figure 13: Image with Text and its Heatmap

7 Expected Outcomes

Review	**Module**	**Input**	**Output**
Review 1	Dataset Collection	-	-
Review 1	Image Feature Extraction	Medical Image	Extracted Features
Review 2	Question Feature Extraction	Question	Extracted Features
Review 2	VQA Model Building	Image + Question	Required Answer
Review 3	Performance Analysis and XAI	Answer + VQA Trained Model	Result of Analysis

8 Conclusion

The dataset for the task of Visual Question Answering is collected and it is analyzed. To extract features that best represents the images, Deep Learning techniques such as CNN, pre-trained VGGNet and VGGNet trained on ImageClef 2019 VQA-Med datset are implemented. The performance of these models are compared in terms of accuracy and visualization(heatmap). The comparison indicates that VGGNet with pre-trained weights perform better in terms of accuracy and also visualization.

References

[1] A. Lubna, S. Kalady and A. Lijiya, *MoBVQA: A Modality based Medical Image Visual Question Answering System*, TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 727-732, doi: 10.1109/TENCON.2019.8929456.

[2] Allaouzi, Imane et al. *An Encoder-Decoder Model for Visual Question Answering in the Medical Domain*, Conference and Labs of the Evaluation Forum (2019).

[3] Al-Sadi, Aisha & Al-Ayyoub, Mahmoud & Jararweh, Yaser & Costen, F.. (2021). *Visual Question Answering in the Medical Domain Based on Deep Learning Approaches: A Comprehensive Study. Pattern Recognition Letters.* 150. 10.1016/j.patrec.2021.07.002.

[4] Sharma, D., Purushotham, S. & Reddy, C.K. *MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain.* Sci Rep 11, 19826 (2021).

[5] Yangyang Zhou, Xin Kang, Fuji Ren. *Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering.* CLEF (Working Notes) 2018.

[6] Knapič S, Malhi A, Saluja R, Främling K. *Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain*. Machine Learning and Knowledge Extraction. 2021; 3(3):740-770.

[7] S. H. P. Abeyagunasekera, Y. Perera, K. Chamara, U. Kaushalya, P. Sumathipala and O. Senaweera, *LISA : Enhance the explainability of medical images unifying current XAI techniques*. 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-9, doi: 10.1109/I2CT54291.2022.9824840.

[8] Lin Z, Zhang D, Tac Q, Shi D, Haffari G, Wu Q, He M, Ge Z. *Medical visual question answering: A survey*. arXiv preprint arXiv:2111.10056. 2021 Nov 19.