Medical Visual Question Answering

Group 42

Nidhin Mohan- B180948CS
Shaaheen A M- B181134CS
Abhinav B Naik- B170297CS

Guided by:

Dr. Saidalavi Kalady
Dr. S. Sheerazuddin

Introduction

Medical Visual Question Answering(Med-VQA)

- Med-VQA is a subtask of VQA, where we perform VQA on medical images. It has a wide variety of uses especially in the field of radiology and also in the diagnosis of tumours.

- For our project we had trained our model using the data available in the SLAKE[12] dataset, which had 642 images and 14,028 question-answer pairs.

- The existing Medical VQA models present in the papers we reviewed had on average about 50-70% accuracies.

Problem Statement

To understand a given medical image, a question relevant to the image and to correctly predict the answer to the question based on the information present in the image, using deep learning techniques.

Literature Survey

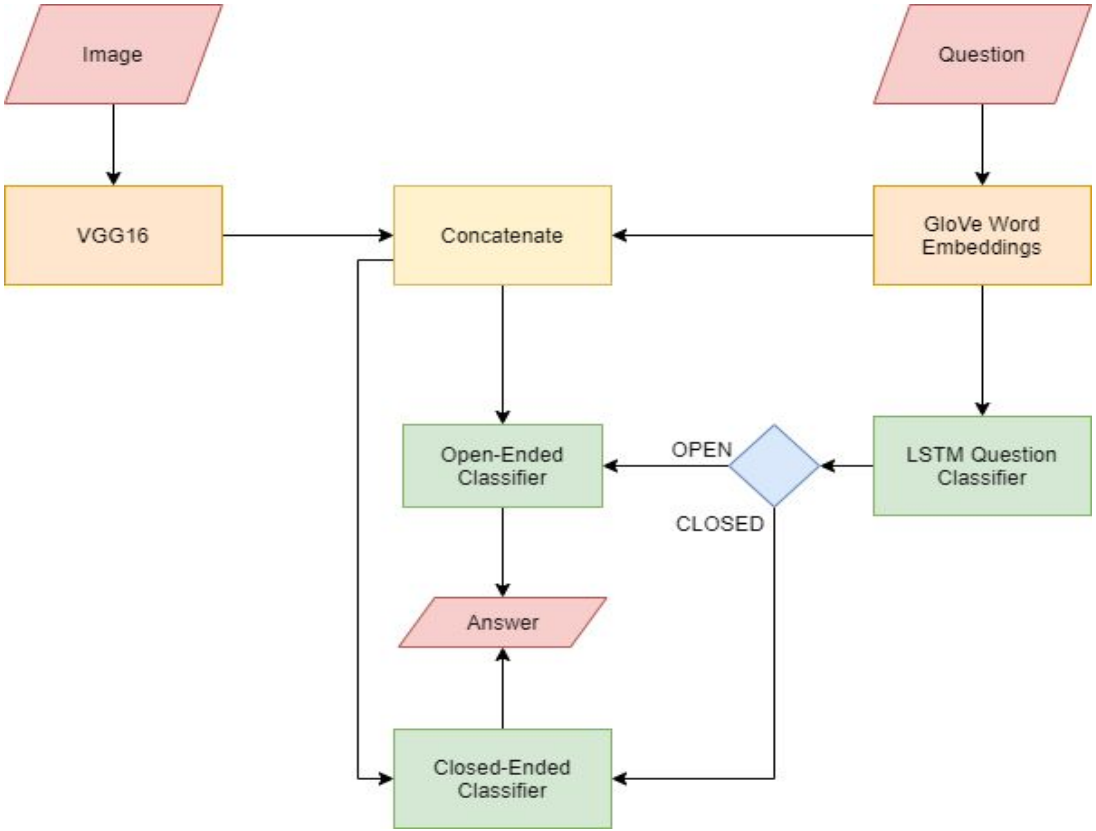
Paper	Methodology	Results
[3] AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping	Used a knowledge inference methodology called Skeleton Based Sentence Mapping.	Accuracy–0.496 BLEU[4]–0.542
[5] Medical Visual Question Answering via Conditional Reasoning	Used Question Conditioned Reasoning module and Task Conditioned Reasoning module to classify questions.	Accuracy-60%
[6] Visual Question Answering in the medical domain based on deep learning approaches: A comprehensive study	The question was classified into four types: organ system, plane, modality and abnormality.	Organ system-75.2% Plane-77.6% Modality-72% Abnormality-18.4%

Paper	Methodology	Results
[7] The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG	Questions were classified into either closed or open ended type and the answer was predicted accordingly.	Accuracy- 0.48, BLEU-0.511
[8] A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering	Used a deep neural network with CNNs to obtain image embeddings.	BLEU score- 0.188. WBSS- 0.209.
[9] Hierarchical deep multi-modal network for medical visual question answering	An SVM-based question segregation module separates the learning path based on open or closed-ended questions.	BLEU-0.0365. WBSS-0.173.

Paper	Methodology	Results
[10] CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering	Questions were classified into five categories: Image features obtained using RestNet152.	Accuracy-0.64. BLEU-0.659. WBSS-0.678.
[11] MMBERT: Multimodal BERT Pretraining for Improved Medical VQA	They used a Multimodal BERT model, pretrained on a VQA Dataset. They used Resnet-152 for Image feature extraction.	Accuracy-0.672. BLEU-0.69.

Proposed Work

Proposed Medical VQA Framework



Obtaining the Necessary Data

- The SLAKE Dataset is used for training our Medical VQA model.
- It contains 9,835 training examples and 2,094 testing examples. Each example is a question-answer pair
- Question-answers are divided into open-ended and closed-ended questions, with the dataset containing both English and Chinese questions
- Questions are of two base types: vqa and kvqa

- The Questions are four types based on their answers:
  - Open-ended single answer
  - Open-ended multiple answer
  - Multiple Choice
  - Closed-ended Yes/No

- For our framework, we used the English open-ended single answer and closed-ended yes/no questions

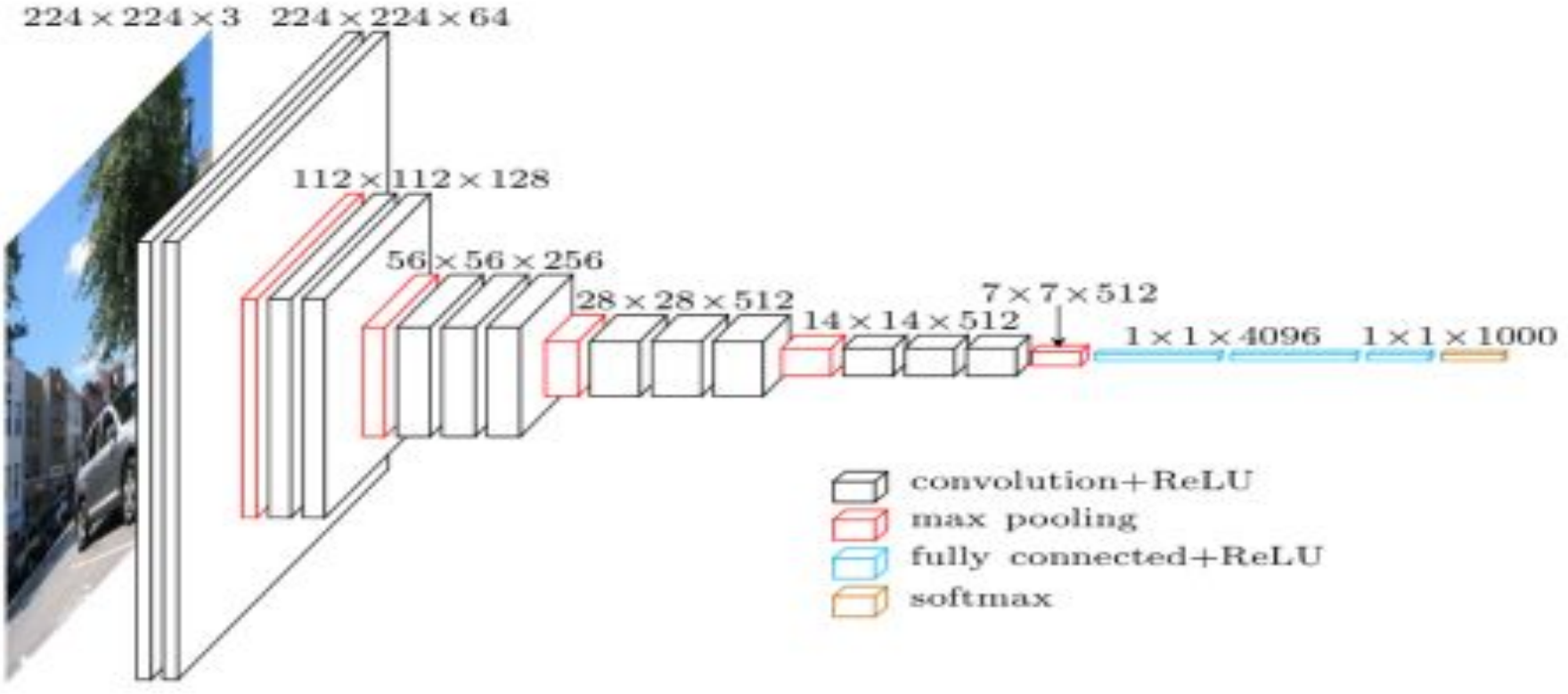
- Training set -> 3949 questions
- Test set -> 831 questions
- Validation set -> 814 questions

Image Feature Extraction

- Model used- **VGG16**[14] pretrained on **ImageNet**[16] database

- Input- **224x224x3** dimensional images
- Output- **7x7x512** feature vector without the final four layers of VGG16

- This feature vector is flattened into a **1x25088** array
- Each image feature is combined into a single array with rows being the flattened image feature



VGG16 model architecture. For our project, we have removed the final three fully connected layers and the softmax layer since we only want the image features.

Question Feature Extraction

- Weights used- **GloVe word embeddings.**
- Input- question as a string.

  Question -tokenizer-> **[1, 5, 4, 17, 95]** -padding-> **[1, 5, 4, 17, 95, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]**

- The padded sequence is passed through an **Embedding** layer which assigns **GloVe word embedding** weights to each word(**1x100** vector) and then flattened.
- Output- **1x1600** dimensional vector.

Question Classification

- Model used- **Sequential** model with a **Bi-LSTM** layer of **50** units and a **Dropout** layer**.**
- Input- **1x1600** dimensional question vector.
- Output- a float value between **0 and 1**.
- The question classification model classifies the question vector into **yes/no** or **open-ended** questions.
- The result of this question classifier determines whether to use open ended or yes/no answer classifier to predict the answers.

Feature Fusion

- Inputs- **1x25088** dimensional image vector and **1x1600** dimensional question vector.
- Output- **1x26688** dimensional fused vector.
- Have performed the simplest feature fusion method, that is concatenating the two features together.
- For each question, the corresponding image feature vector is retrieved and concatenated to obtain a **1x26688** feature vector which is then passed to the corresponding classification model.

Classification

- Two separate classifiers for obtaining the final answers :

  [1] **Softmax classifier :** classifies open-ended questions into 1 of 124 answer classes.

  [2] **Sigmoid classifier :** classifies closed-ended questions into either Yes or No.

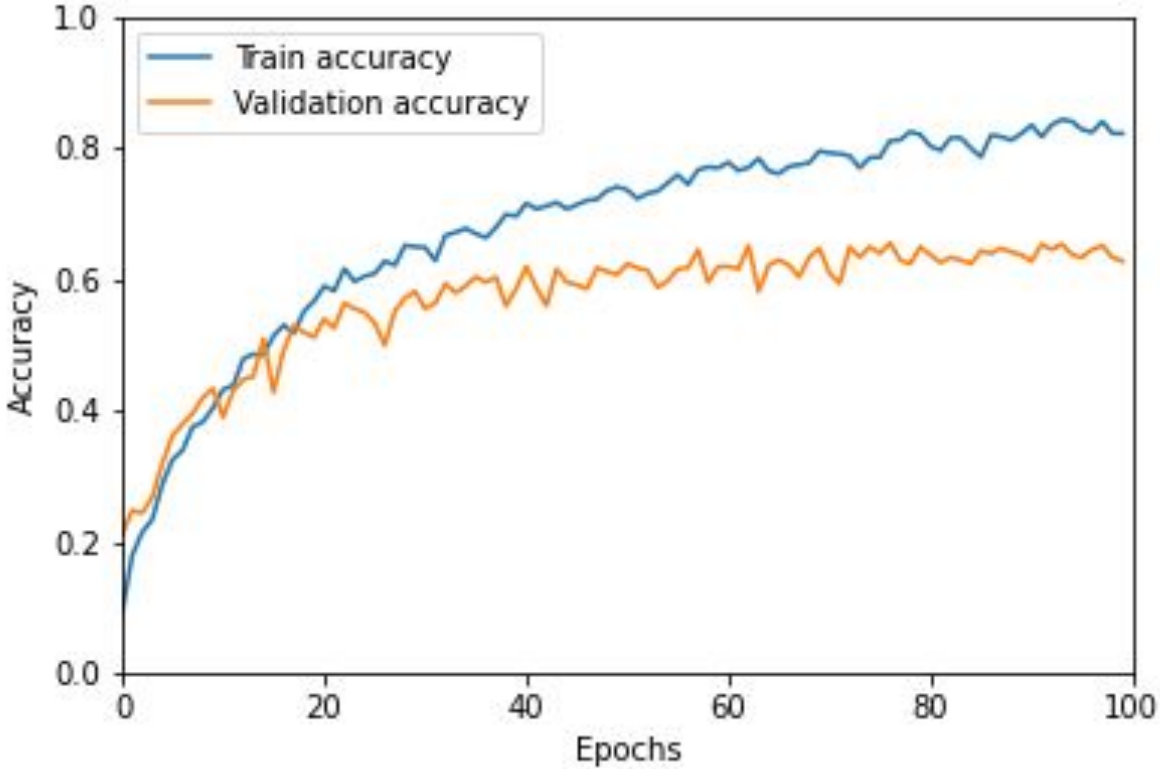
- Image-question fused vector is passed to the corresponding classifier based on the output of the question classifier.

Experimental Results

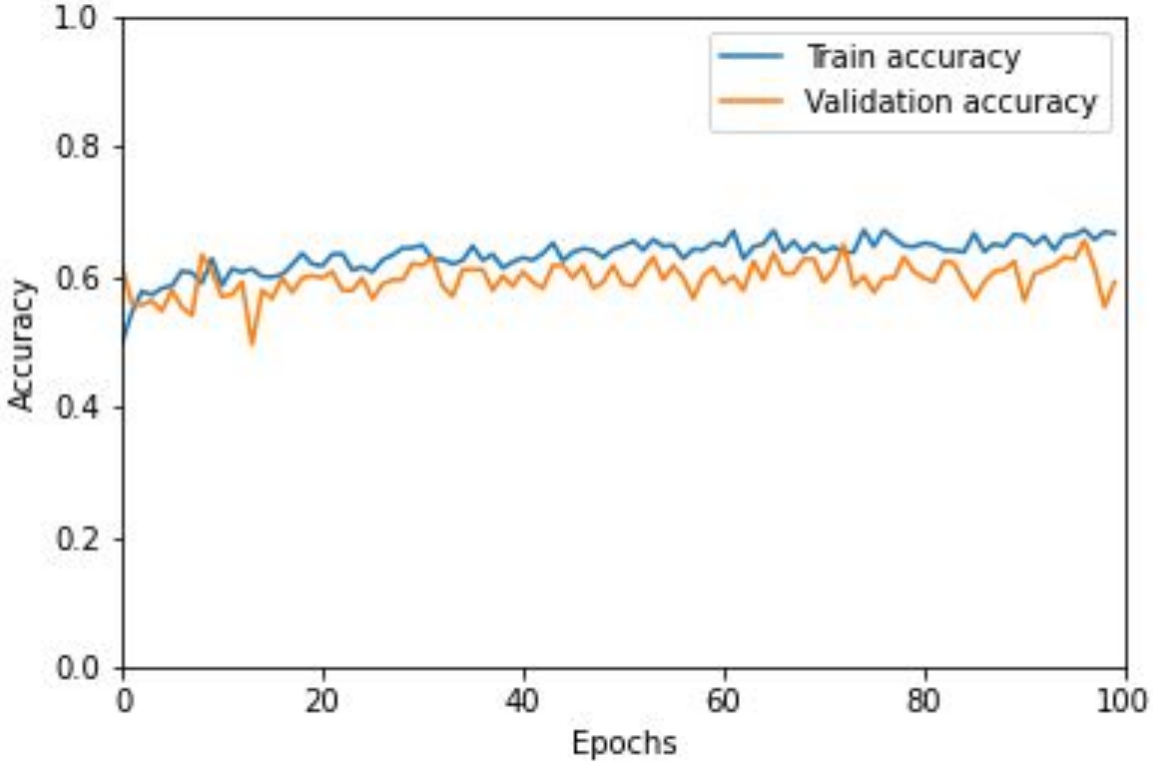
- Question Classifier gave an accuracy of 100%.

- Open ended image-question classifier model gave an accuracy of 62.91% (515 open-ended image-question pairs).

- Closed ended image-question classifier model gave an accuracy of 61.08% (316 closed-ended image-question pairs).



Training and Validation Accuracies against epochs for Open-Ended Question Answer Classifier



Training and Validation Accuracies against epochs for Closed-Ended Question Answer Classifier

What is the abnormality in this image?
Lung Cancer

Is this MRI?
Yes

How many lungs are in this image?
0

Some examples when manual testing that gave correct answers

Where is the liver in this image?
Top

What diseases are in this image?
Lung

Some examples when manual testing that gave incorrect answers

Conclusion

- Model has given good accuracies on the training, testing and validation datasets.

- Questions could be segregated based on the content and appropriate information could be retrieved from the images to answer the questions.

- Use of better feature fusion methods can improve the accuracy of the model.

References

[1] Jeffrey Pennington, Richard Socher and Christopher D. Manning. ”GloVe: Global Vectors for Word Representation”.

[2] Jin-Hwa Kim, Jaehyun Jun and Byoung-Tak Zhang. ”Bilinear Attention Networks”.

[3] Zhibin Liao, Qi Wu, Chunhua Shen, Anton van den Hengel and Johan W. Verjans. “AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering.”

[4] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. ”BLEU: a Method for Automatic Evaluation of Machine Translation”.

[5] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical Visual Question Answering via Conditional Reasoning.

[6] Aisha Al-Sadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Fumie Costen, “Visual question answering in the medical domain based on deep learning approaches: A comprehensive study”

[7] Aisha Al-Sadi, Hana AL-Theiabat, Mahmoud Al-Ayyoub. (2020). The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG.

[8] R. Ambati and C. Reddy Dudyala, ”A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering”.

[9] Deepak Gupta, Swati Suman, Asif Ekbal. (2020). Hierarchical Deep Multi-modal Network for Medical Visual Question Answering.

[10] F. Ren and Y. Zhou, ”CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering”.

[11] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar and C. Jawahar, ”MMBERT: Multimodal BERT Pretraining for Improved Medical VQA,”.

[12] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, Xiao-Ming Wu. ”SLAKE: A Semantically-Labelled Knowledge-Enhanced Dataset For Medical Visual Question Answering”.

[13] Chelsea Finn, Pieter Abbeel, Sergey Levine. ”Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”.

[14] Karen Simonyan, Andrew Zisserman. ”Very Deep Convolutional Networks for Large-Scale Image Recognition”.

[15] Tuong Do, Binh X. Nguyen, Erman Tjiputra, Minh Tran, Quang D. Tran, Anh Nguyen. ”Multiple Meta-model Quantifying for Medical Visual Question Answering”.

[16] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, ”ImageNet: A large-scale hierarchical image database”.

Thank You