Medical VQA

Using ML to answer questions about medical imaging data.

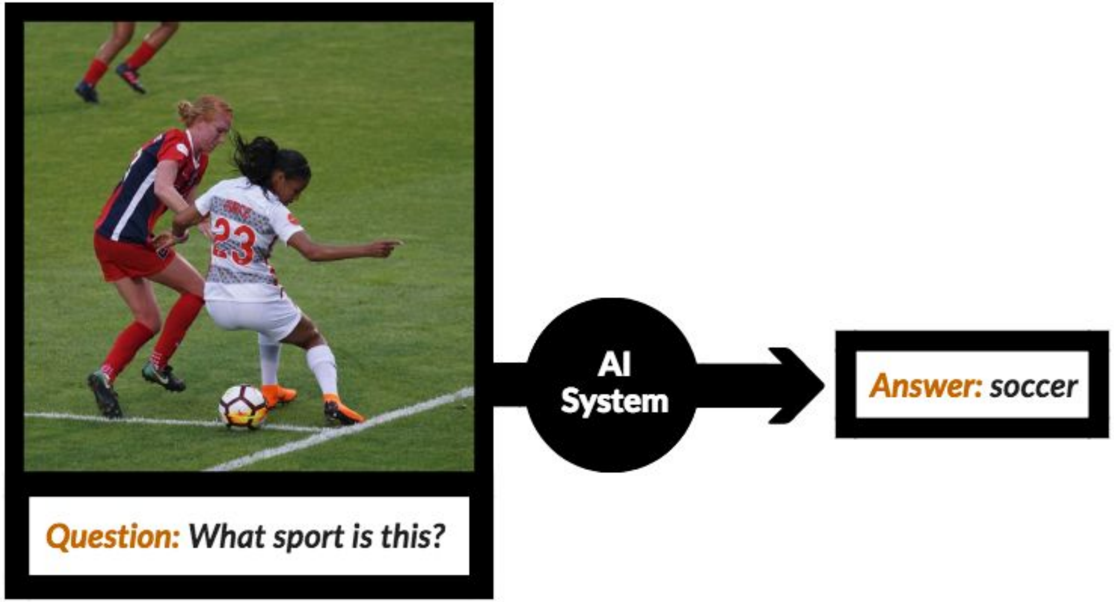
Contents

Introduction

We’ll discuss the general
concept of VQA.

VQA: Visual Question Answering

In **VQA** the **context** is inferred from a given image



VQA: Visual Question Answering

- **VQA** is done using datasets made specifically for the task

- The first VQA dataset was created in **2015**

What color are her eyes?
What is the mustache made of?

VQA Applications

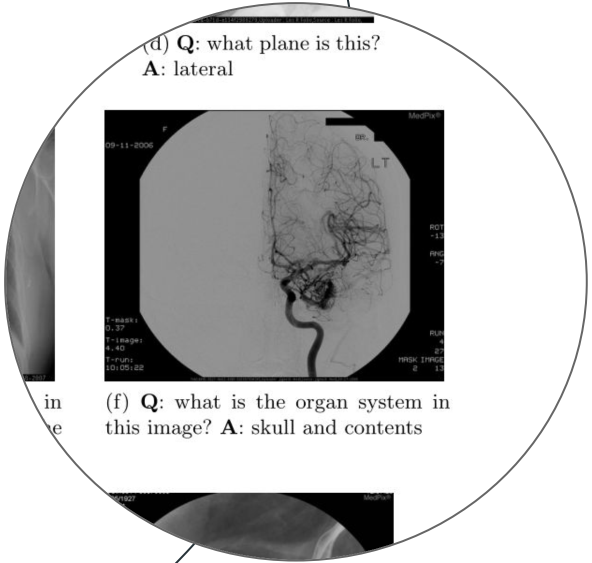
smarthomes

Education

Helping Visually impaired people

Problem Statement

Build a Diagnostic Decision Support System (**DDSS**) that could interact with medical practitioners or students.



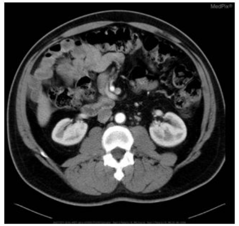
Dataset Description

The dataset was created by a research team from **Lister Hill National Center for Biomedical Communications**.

Source

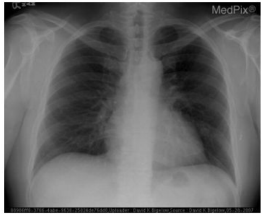
images are from **MedPix®**, an open-access radiology archive of case reports.

Sample QAs

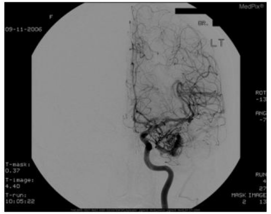


(c) **Q**: is this a contrast or non-contrast ct? **A**: contrast

(d) **Q**: what plane is this? **A**: lateral



(e) **Q**: what abnormality is seen in the image? **A**:nodular opacity on the left#metastastic melanoma



(f) **Q**: what is the organ system in this image? **A**: skull and contents

315 Images

3,515 Questions

Mean of 7 words per question, max 21 words.

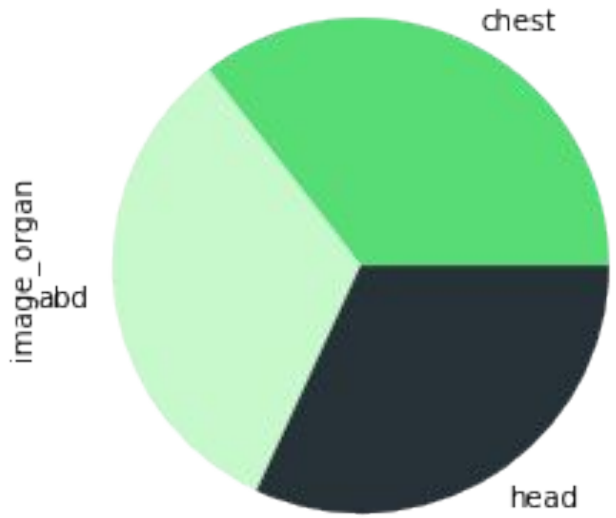
476 Answers

Final Dataset

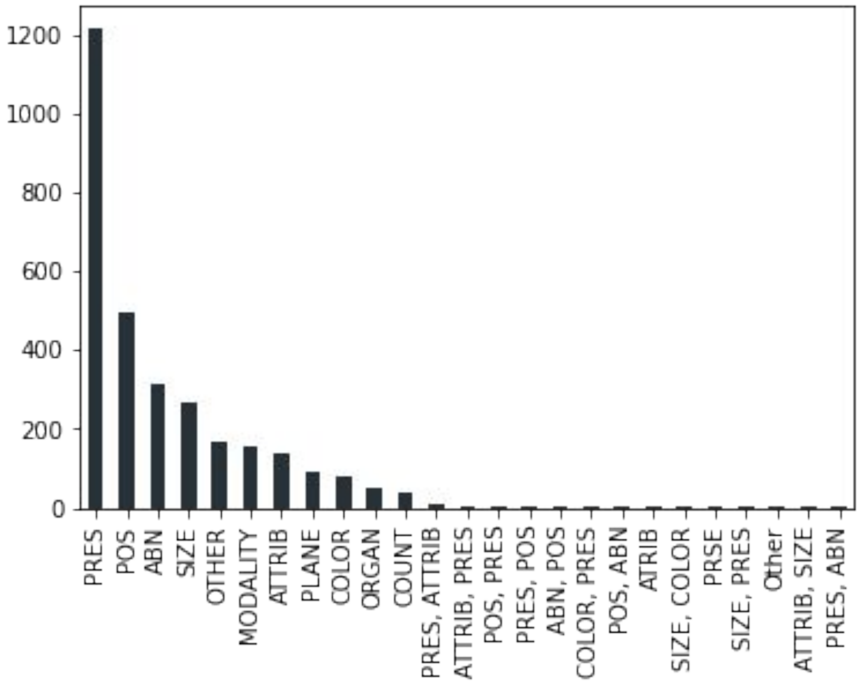
qid	Question id.
image_name	Image id.
image_organ	Name of organ system: `'head', 'chest', 'abd'`
answer	The volunteers answer
answer_type	Closed or Open-ended
question_type	Abnormality, location, presence, count, size..etc
question	The full-text of the question
phrase_type	Free-form or not

EDA

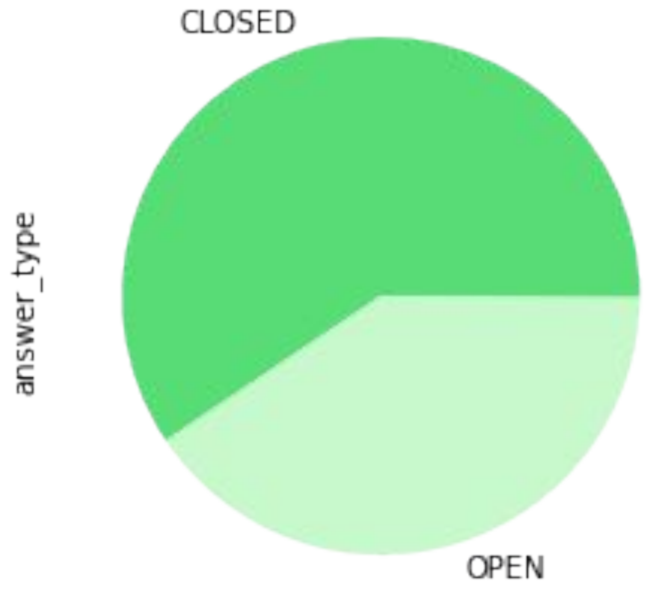
EDA: Image Organ distribution



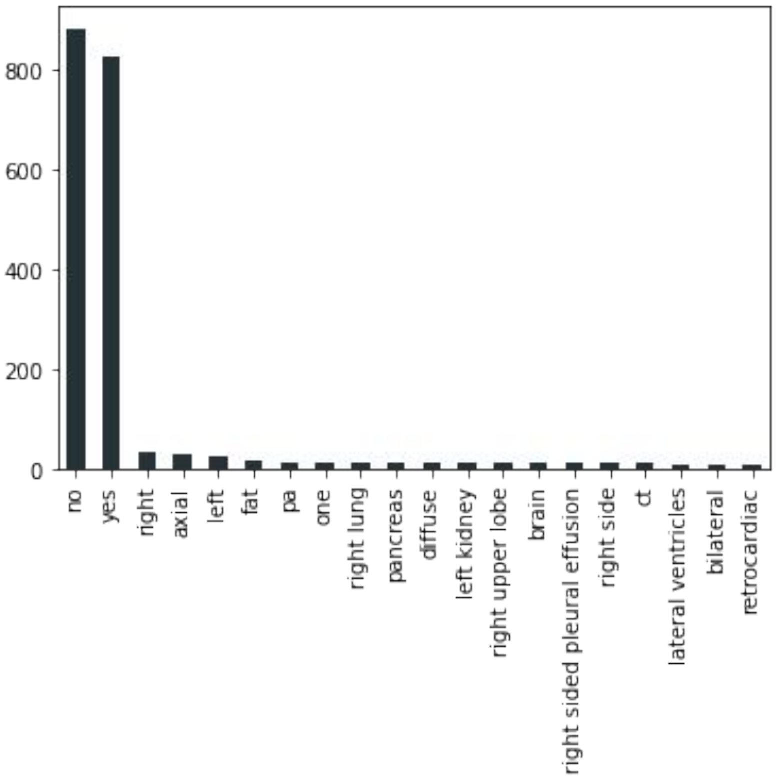
EDA: Question Category



EDA: Answer Category distribution



EDA: Answer distribution



VQA Solution Approach

Steps To Solve VQA problems

1

Image Feature extraction

using pre-trained CNN

2

Question Feature extraction

Usually using LSTM

3

Conctination

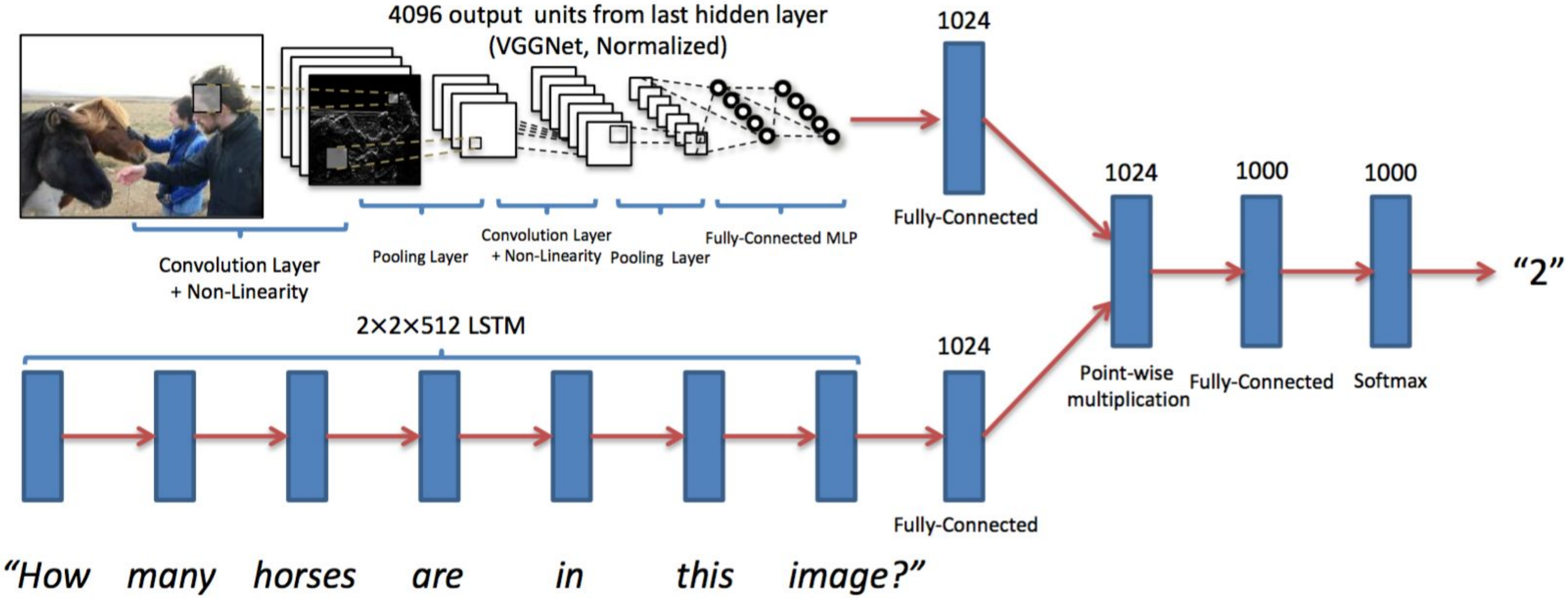
Using pointwise multiplication

4

Predict Answer

Softmax

Baseline Architecture



Question Model

Preprocessing

Regular text preprocessing (lowercase, remove punctionations..etc), followed by tokenization

LSTM

To get feature vector

Word-embedding

After tokenization questions are embedded Using word2vec

Question Model

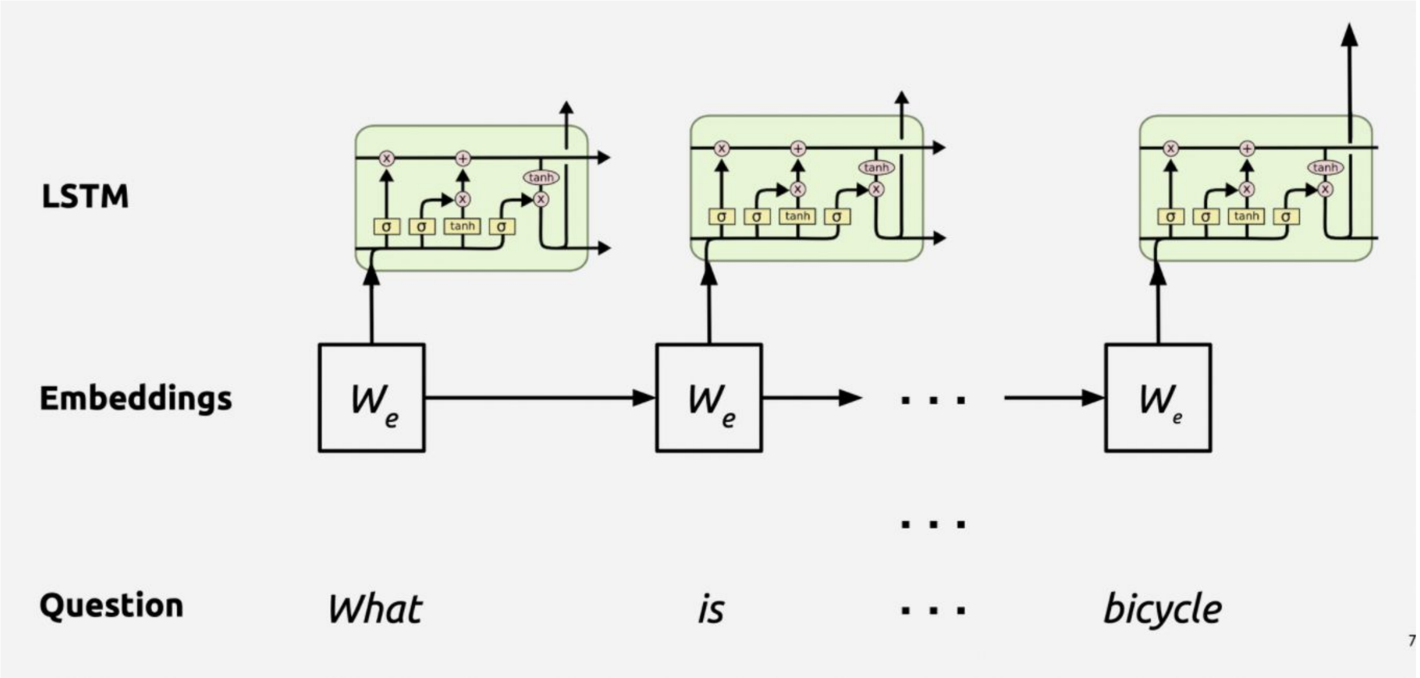


Image Model : using Vgg16

1

Rescale Images
to 448 x 448

2

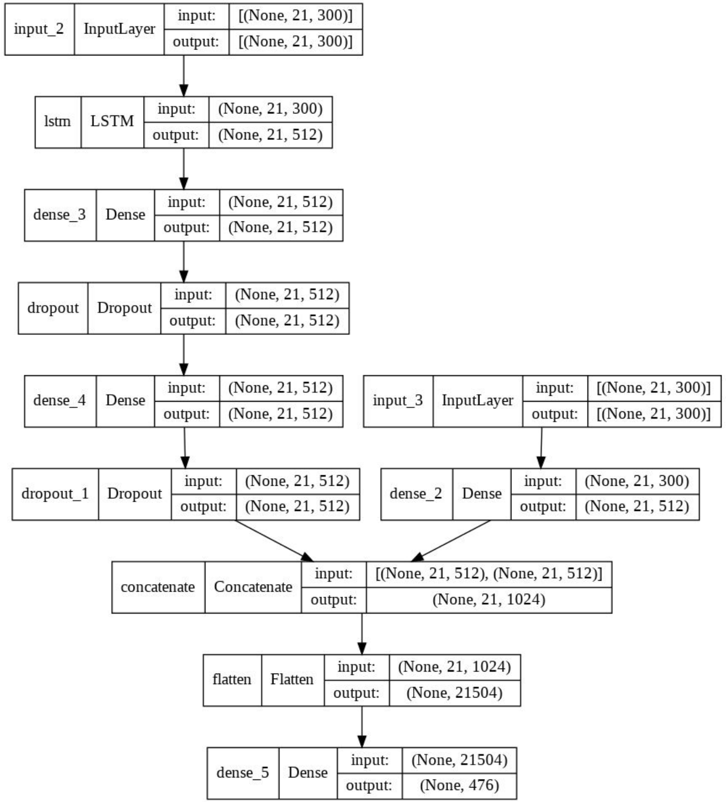
Get the features
from the last Vgg16
pooling layer

3

Transform Vectors
to be the same
shape of the
question vector

Model

Model plot



- Used tanh activation function.
- Used adam optimizer, and a learning rate of 0.001, 10 epochs, batch size of 16.

Results

Results

	Accuracy	Loss
Type	Simple accuracy	Sparse Categorical Cross-entropy
Train	97%	0.4599
Validation	43%	5.8672

Sample inference (Close-ended)

Question: Is there evidence of an aortic aneurysm?
answer: Actual: yes predicted: yes

Question: Does this patient have evidence of pneumoperitoneum?
answer: Actual: No predicted: no

Sample inference (Open-ended)

```
Question: How is the patient oriented?
answer: Actual: Posterior-Anterior predicted: ultrasound
-------------------------------------------------
```

Original Paper Results

As might be expected, the models that best predicted VQA-RAD test set are trained on VQA-RAD training alone or partially. For closed-ended questions, MCB_RAD had slightly better simple accuracy of 60.6%, while SAN_RAD had a slightly better mean accuracy at 54.6%. This performance is comparable to the MCB and SAN models trained and tested on VQA1.0 dataset which had reported 64.2%[16] and 58.9%[17] accuracies for open and closed ended public domain questions. Predicting open-ended VQA-RAD questions, the models performed much lower with the MCB_RAD scoring the best at 25.4% simple accuracy and 19.3% mean accuracy. The contrast between open and closed-ended questions suggests that these models are currently still guessing and may require more data. Potential improvements can come from learning medical terminology and jargon, greater quantity and diversity of questions, and extracting image features from radiological images as opposed to public domain.

Future Work

Future Work

1

Data

I'll Try to use the full dataset or the newer versions.

2

Architecture

I'll try attention based architectures.

Future Work

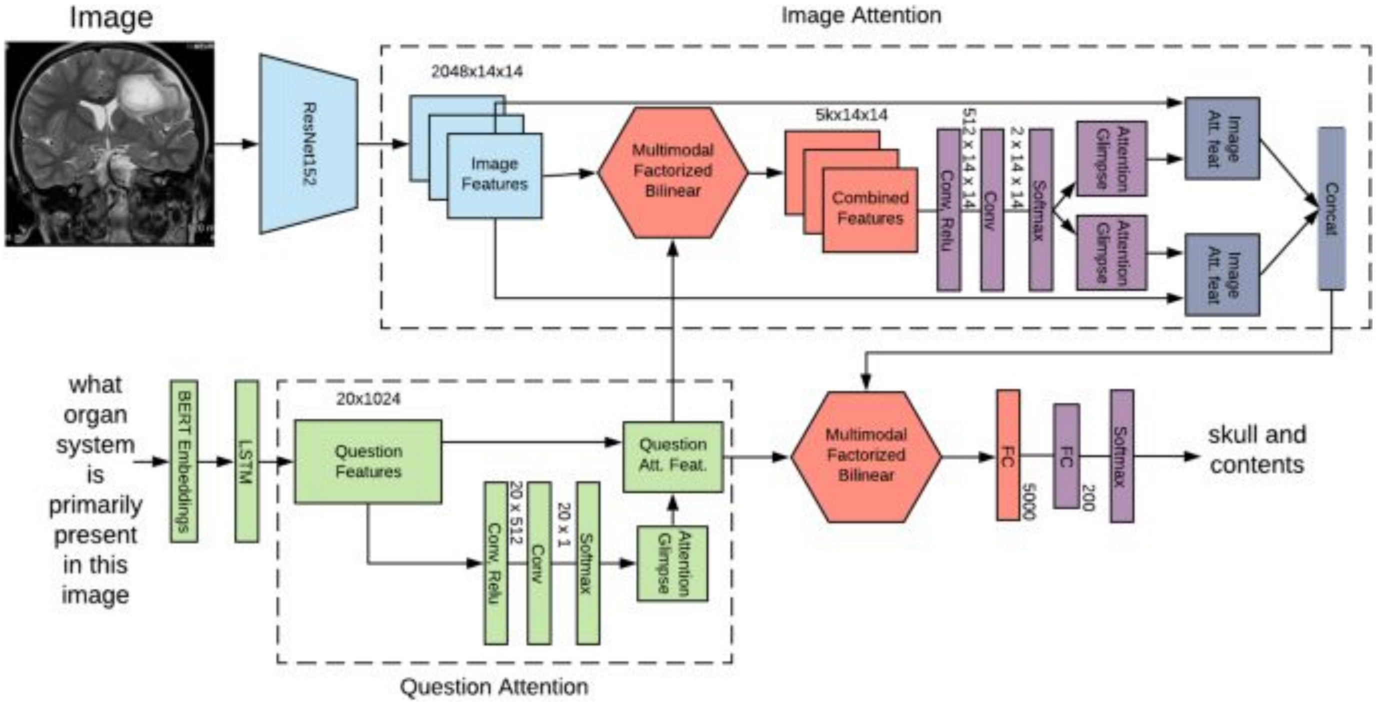


Image
Image Attention
ResNet152
2048x14x14
Image Features
Multimodal Factorized Bilinear
5kx14x14
Combined Features
Conv, Relu
512 x 14 x 14
Conv
2 x 14 x 14
Softmax
Attention Glimpse
Attention Glimpse
Image Att. feat
Image Att. feat
Concat
what organ system is primarily present in this image
BERT Embeddings
LSTM
20x1024
Question Features
Question Att. Feat.
Conv, Relu
20 x 512
Conv
20 x 1
Softmax
Attention Glimpse
Question Attention
Multimodal Factorized Bilinear
FC
5000
FC
200
Softmax
skull and contents

Conclusion

Using the baseline architecture with this dataset leads to overfitting, the dataset is small and imbalanced for such complicated task. The model performs better on close-ended questions, but it performs badly on open-ended ones, this could be improved by using more data and trying different architectures.

Thanks!

any questions?