



## Introduction

- Visual Question Answering (VQA) in the medical domain aims to answer a clinical question presented with a medical image. The system could support clinical education, clinical decision, and patient education.
- For doctors, It helps to interpret complex clinical images and make more accurate clinical decisions.
- For patients, it leads to better understand their health condition.

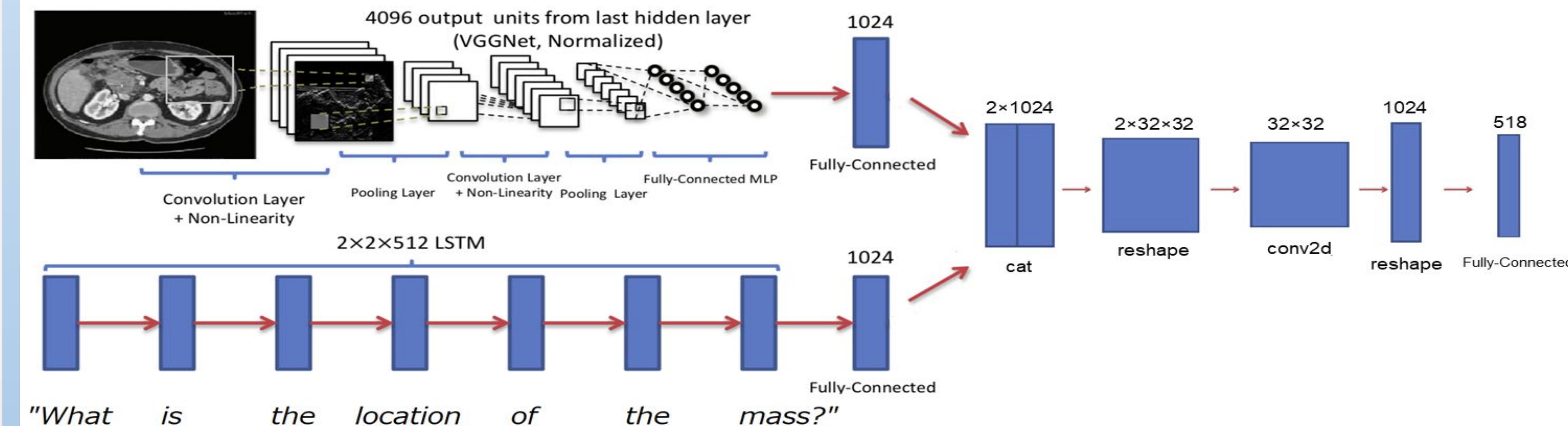
## Data

- VQA-RAD dataset that contains clinically generated visual questions and answers about radiology images.
- It contains 314 images, 2248 objects, 1019 kinds of questions, 517 kinds of answers.
- 42% is open-ended and 58% is close-ended

## Method

- Image Encoder: use VGG19 to catch image features [1, 1024, 1]
- Question Encoder: use LSTM to catch question features [1, 1024, 1]
- Fusion: Concatenate and reshape the two features to [2, 32, 32], use Conv2D to fuse them to [1, 32, 32], then reshape it to [1, 1024, 1]
- Fully Connected: Use fully connected layer to calculate the possibility of each answer.

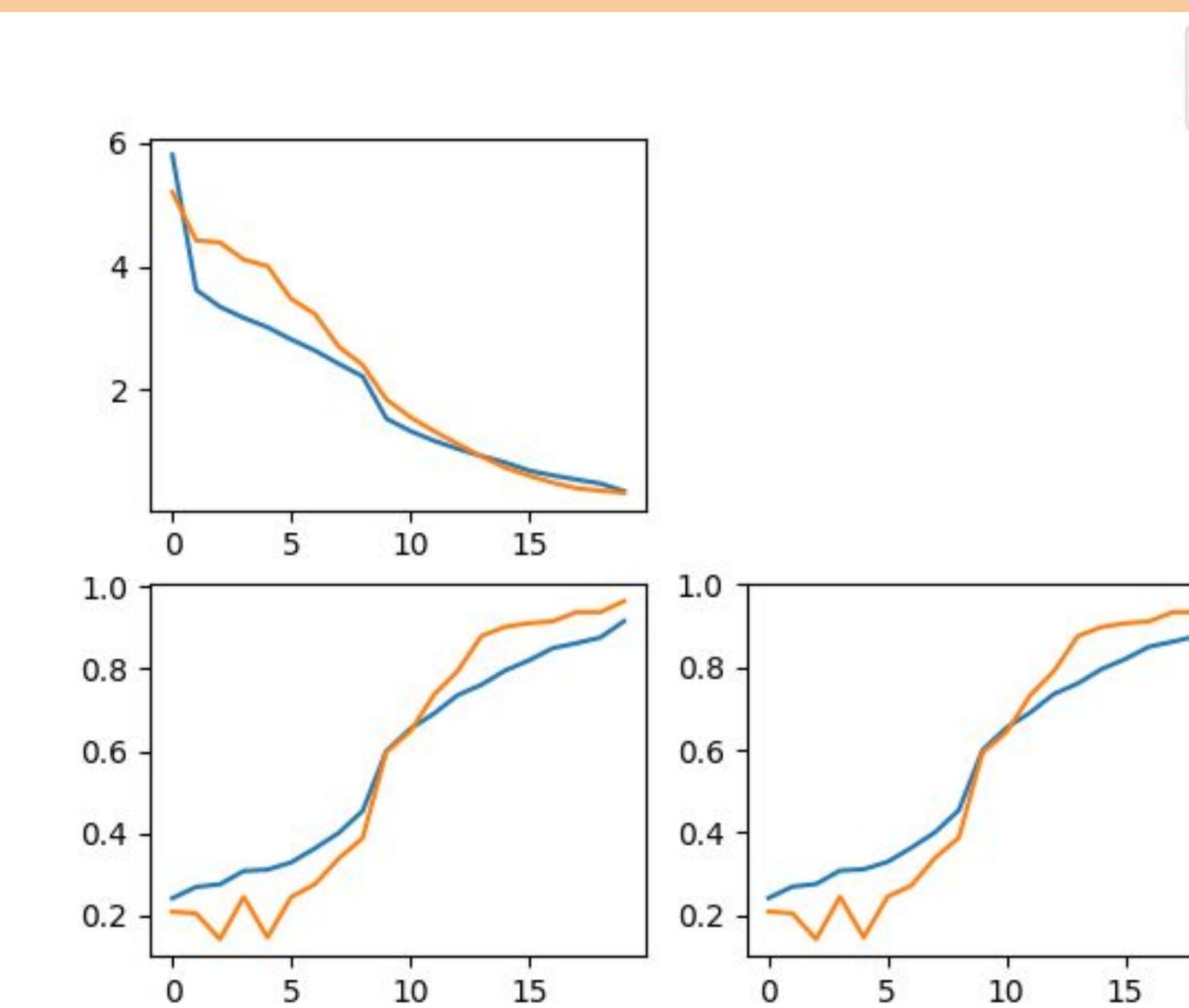
## Model



## Prediction



## Accuracy



## Answer Field



question:	What organ system is pictured ?
answers	possibility 8.830191
the brain	8.558833
brain	7.821183

question:	What type of imaging is this ?
answers	possibility 4.9842787
mri	4.8290596
xray	4.6724877

question:	Is this an MRI ?
answers	possibility 3.0981603
yes	2.8502588
no	-2.938237

## Result & Analysis

- The best result is the 10th model with mean and standard deviation based on the dataset, and batch size 1
- Small dataset causes overfitting quickly, since the train accuracy reached almost 100% in the end.
- The model can predict simple open-ended questions accurately, but performs not well on close-ended questions.
- The inaccurate results are caused by lack of questions, since different organs and symptoms can be combined into a lot of close-ended questions. It is hard to train all of them.

## Conclusion & future work

- LSTM still has limitation on question encoding.
- We need to import more tools (NLP modules) or a more efficient model (attention model) to help understand questions.
- Data augmentation is also one of possible solutions, but may produce invalid images or break the balance of yes/no answers.
- Transfer Learning may also give us a better prediction.