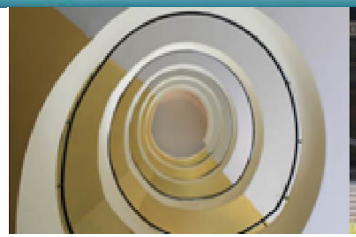
Project Work :

Final ISA(Review 4)

Project Title : Image Captioning for Chest X-Ray images.
Project ID : PW20PRD01
Project Guide : Prof. P. Ramadevi
Project Team : 01FB16ECS482 Krishna Khurana
01FB16ECS488 Manmath Sahoo
01FB16ECS490 Shrey Jain

PES UNIVERSITY






Problem Statement

- Understanding medical images is challenging. It requires medical professionals with certain skills to understand the radiology reports.
- Medical imaging are error-prone. Human generated report can have some errors, which can impact proper treatment of patients.
- The process of writing the impression statements is time-consuming and highly repetitive with the dictation of the findings. This suggests a crucial need to automate the radiology impression generation process

User Profile

Distinct user facing the problems:

- Radiologists and pathologists who examine various medical images daily like, PET/CT scans and write findings and reports.

- Experienced physicians and other doctors spend a lot of time analysing reports, which is tedious and time consuming.

- Reports generated by the inexperienced doctors can be error prone.

Literature Survey

Title	A survey on Biomedical Image Captioning
Authors	Vasiliki Kougia, John Pavlopoulos, Ion Androutsopoulos Department of Informatice, Athens University Of Economics and Business,Greece
Abstract	This paper is the first survey of image captioning in the medical domain. Additionally it also gives us knowledge of the datasets, the evaluation measures. This paper also suggest two baselines, a weak and a stronger one.
Drawbacks	The paper fails to extent the difference between generic image captioning and bio-medical image captioning and does not have a view of the real-life needs and the degree to which the methods are aligned with their needs.

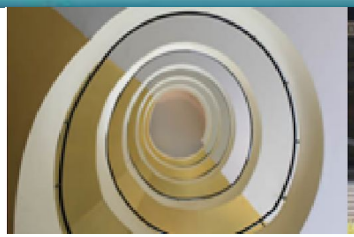
Literature Survey

Title	Learning to Summarize Radiology Findings (2018) https://nlp.stanford.edu/pubs/zhang2018radsum.pdf
Author	Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, Curtis P. Langlotz Stanford University Stanford, CA 94305
Abstract	Propose a solution to generate automate the generation of radiology impressions with neural sequence-to-sequence learning model. A customized neural model is designed to encode the input information and then its output is used for the decoding process. The dataset used is the original dataset from the hospitals.
Drawbacks	The system summary is not accurate for every body parts. According to paper, it works fine for “knees” but the report degrades for “chests”.Most of the errors in the report misses critical information, which changes the overall result of the report.

Literature Survey

Title	From Chest X-rays to Radiology Reports: A Multimodal Machine Learning Approach (2019)
Author	Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shan, Len Hamey Carnegie Mellon University
Abstract	The paper proposes a model based on multimodal machine learning that can automatically generate radiology reports for medial images.It further claims that the previous CNN-RNN based text-generation methods used LSTMs that are relatively shallow in depth.a multistacked CNN-RNN model has a potential to generate semantically rich reports for chest X-ray images.
Drawbacks	The model faced problem generating result for abnormalities. It was able to generate some good results for normal chest x-rays.

PES UNIVERSITY

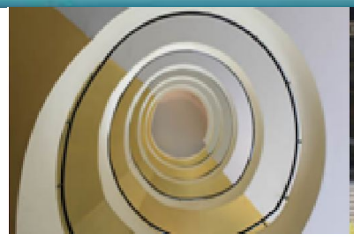




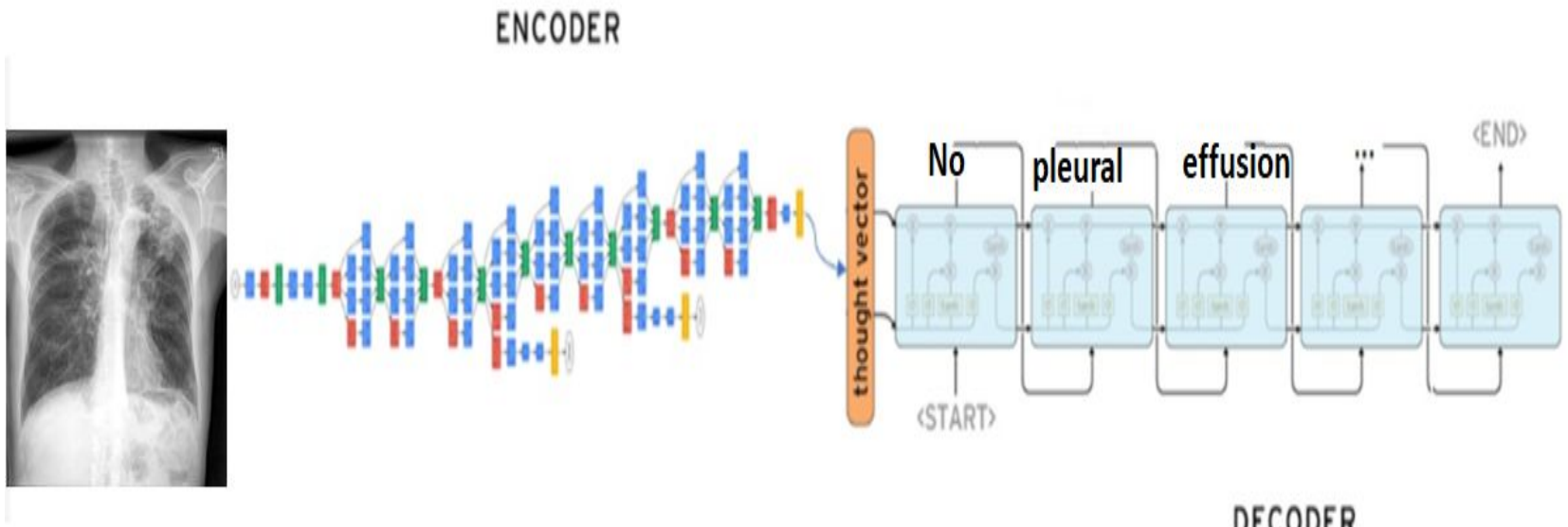

Proposed Solution

- Encoder-Decoder architecture to annotate medical images.
- The input to the encoder is the medical image which we have to caption.
- The encoder consist of the Inception V3 model, in which we have removed the last layer of the model to encode the X-ray image.
- The encoded image from the CNN layer is used as the input to the RNN model, which consist of the Attention mechanism.
- The decoder will provide us with the caption for the images.

PES UNIVERSITY





Why Your Solution is Better?

- We have applied Transfer Learning
- We have used Attention mechanism so that it enables the model to remember all the words in the input and focus on specific words when formulating a response.
- We have also used Teacher Forcing which allows the model to converge faster.

Technologies / Methodologies

Technologies:

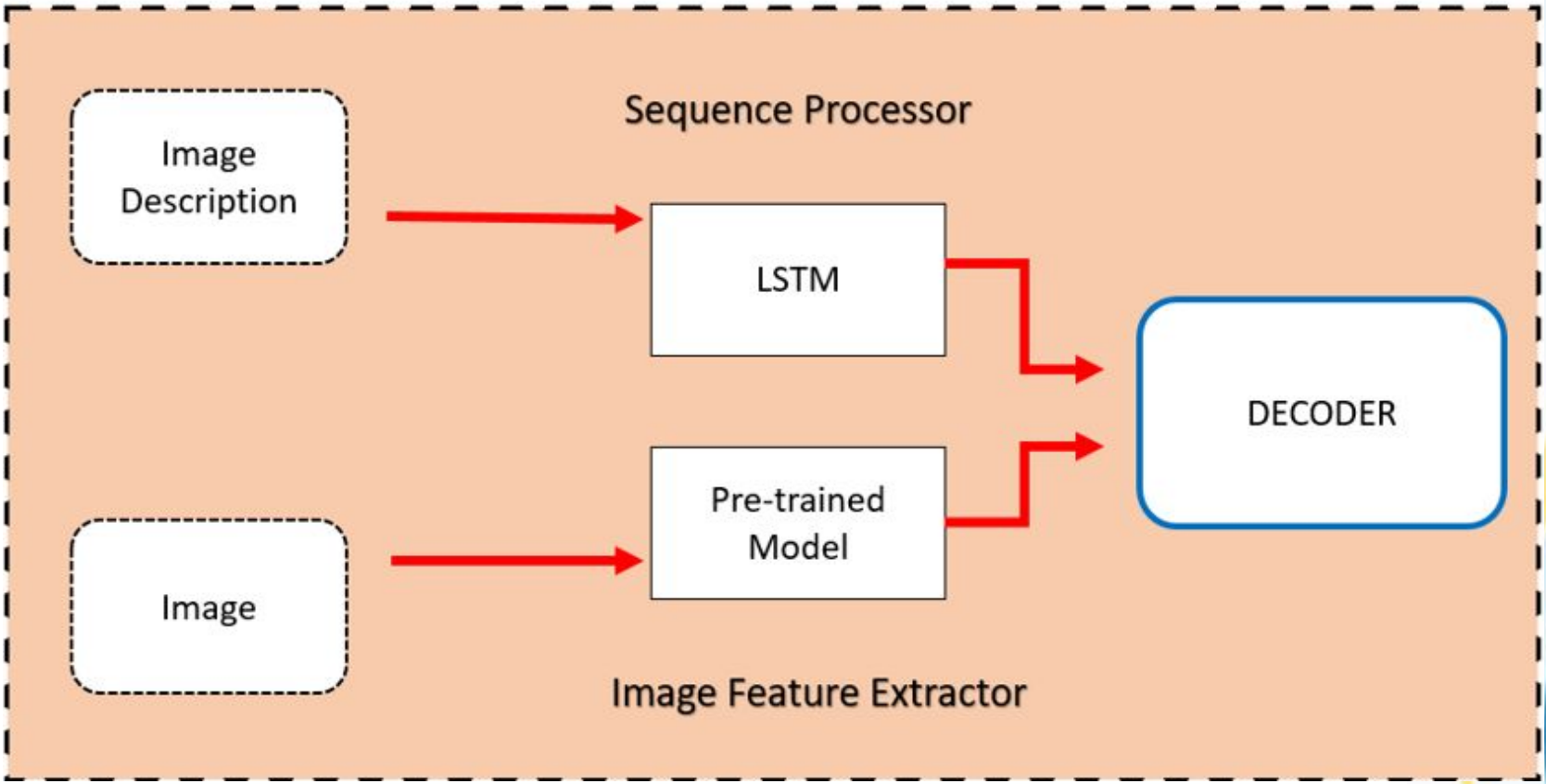
- Python 3 :- Keras with TensorFlow as backend.
- Flask :- For serving the backend i.e. the machine learning core.
- Html,Css,Javascript :- For UI related work.

Dependencies and Risks

- Various modules which run on previous models of keras, can cause verison problems.
- The application should be supported by various web browsers.
- High computation power is needed to train the model.
- The dataset cannot be made public, since it is of medical use. This limits the growth of new technology.
- Only certified radiologists will be given the access to use the product to generate the caption.

System Architecture

Image Captioning Model Architecture

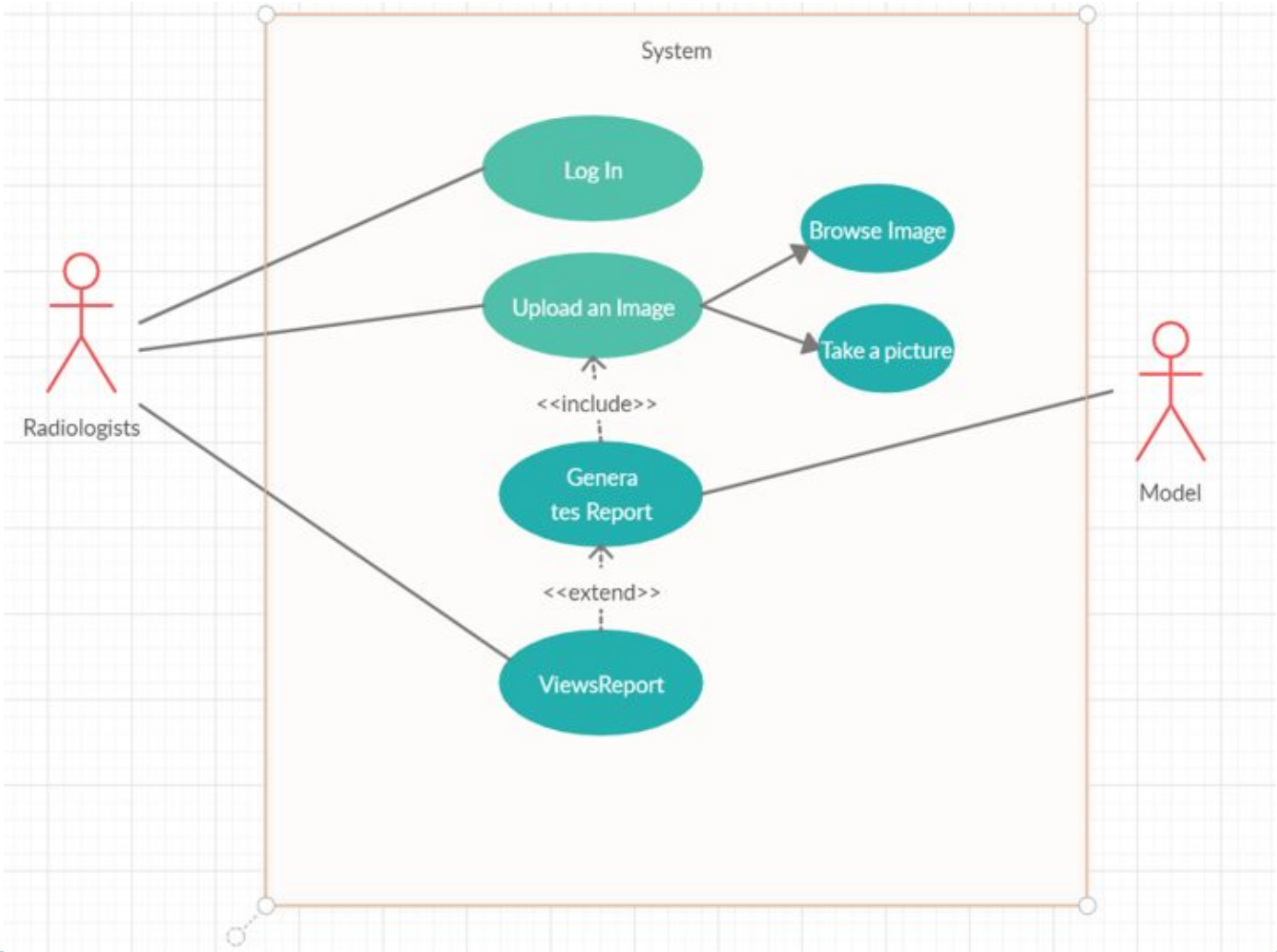


System Architecture

- Encoder-Decoder used for Image Captioning.
- A pre-trained CNN model will be used to extract the features of an image in a dense representation.
- The image description will be converted to n-dimensional vector using pre-trained glove vectors.
- Both the Sequences(description) and Image Features(image) is then merged and fed into the Decoder.

PES UNIVERSITY

UI/ Use Case



Modules

Our Project includes 6 Major Modules/Components listed as below:

1) Download and Prepare IU-X Ray Dataset:

The IU X-Ray dataset comes with images, a txt file having image names, and a caption file containing captions in the same order. The dataset was split into (90-10) train-test split . We load the description, clean it, drop stop words from it and then store it against an image name in a dictionary. We then use these cleaned description to make a vocabulary of the words.

Modules (continued…)

2) Preprocess and tokenize the captions:

The dataset contains multiple descriptions for each photograph and the text of the descriptions requires some minimal cleaning. We clean the text in order to reduce the size of the vocabulary of words we will need to work with:

3) Preprocess the images using InceptionV3

We used InceptionV3 (which is pre trained on Imagenet) to classify each image. We extracted features from the last convolutional layer.First, we resize the image to 299px by 299px. Preprocess the images using the preprocess_input method to normalize the image so that it contains pixels in the range of -1 to 1, which matches the format of the images used to train InceptionV3.

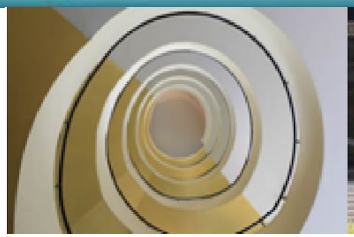
Modules (continued…)

4)Initialize InceptionV3 and load the pretrained Imagenet weights:

We need to convert every image into a fixed sized vector which can then be fed as input to the neural network. For this purpose, we opt for **transfer learning** by using the InceptionV3 model (Convolutional Neural Network) created by Google Research.

5)Encoder

The Convolutional Neural Network(CNN) can be thought of as an encoder. The input image is given to CNN to extract the features. The last hidden state of the CNN is connected to the Decoder.





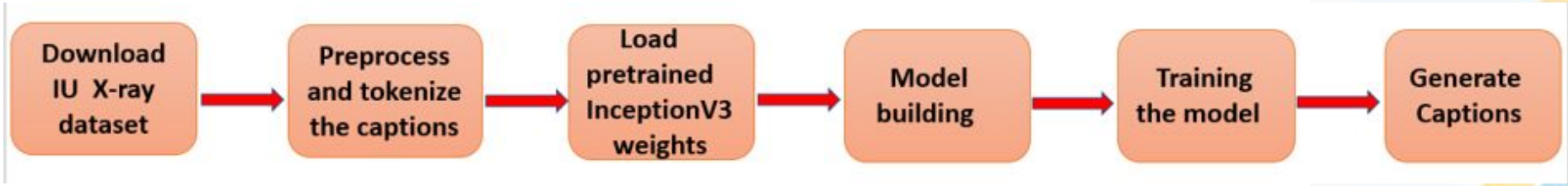

6) Decoder

The Decoder is a Recurrent Neural Network(RNN) which does language modelling up to the word level. The first time step receives the encoded output from the encoder and also the <START> vector.

7) Training

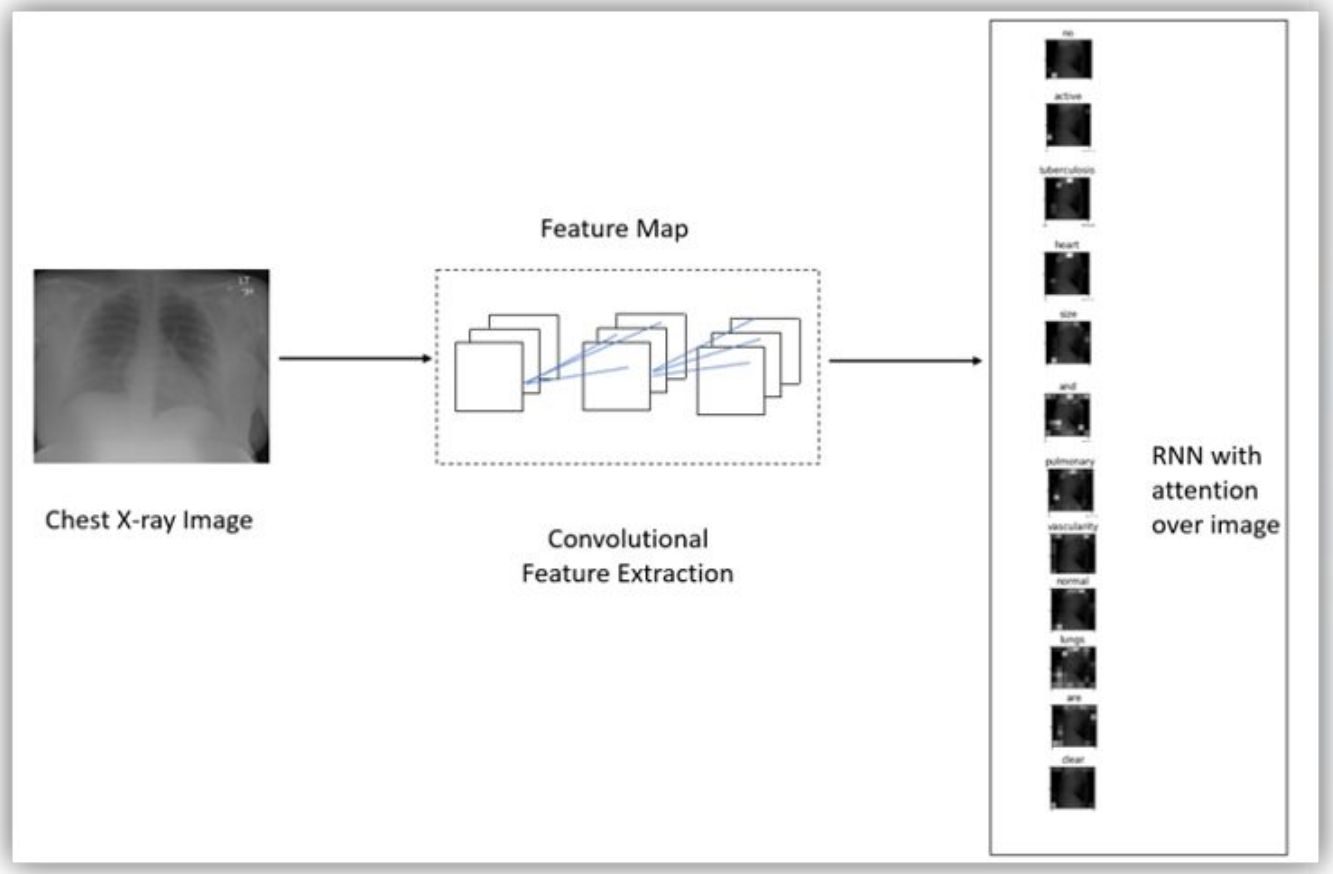
We trained the model progressively for about 80 epochs. And we have a loss of 0.027

Design Approach



Design Approach

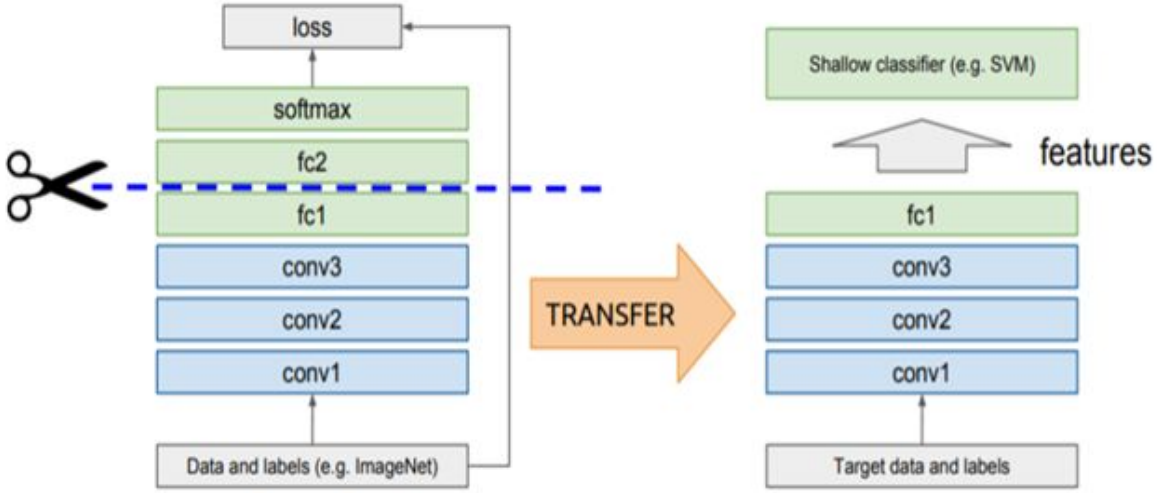
Attention Mechanism



Design Approach

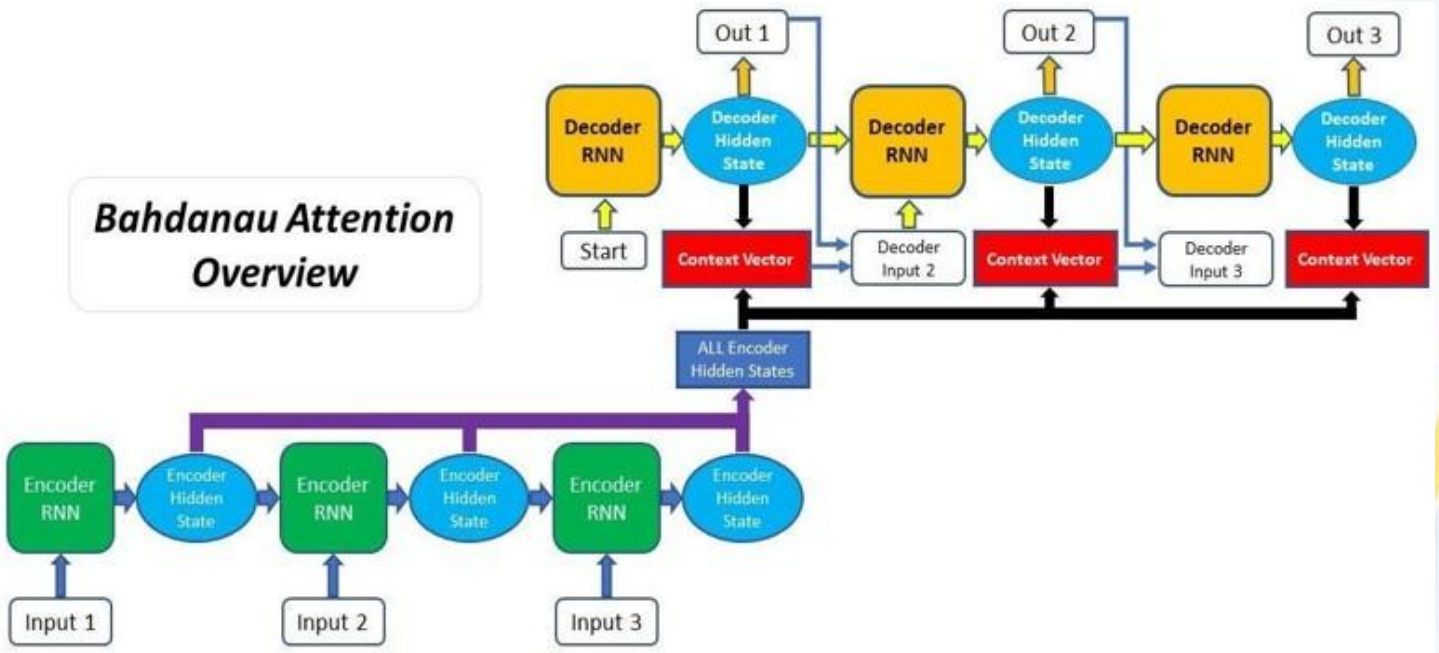
TRANSFER LEARNING

Deep learning systems and models are layered architectures that learn different features at different layers These layers are then finally connected to a last layer to get the final output. This layered architecture allows us to utilize a pre-trained network (such as Inception V3) without its final layer as a fixed feature extractor for our task.



Design Approach

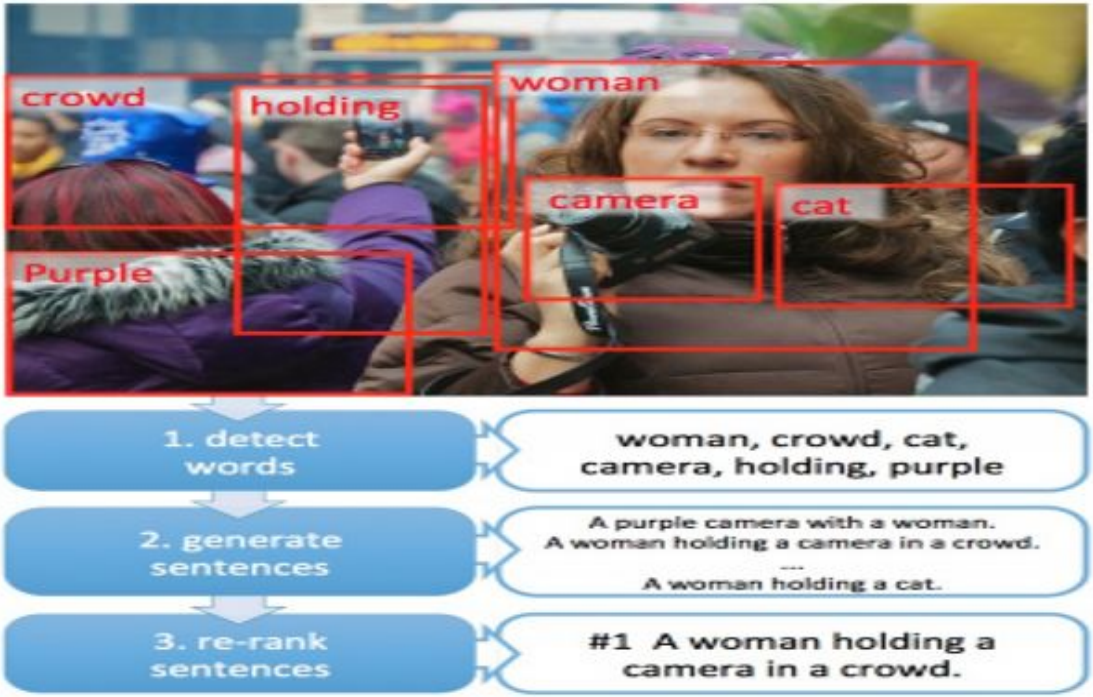
The Attention mechanism has revolutionised the way we create NLP models and is currently a standard fixture in most state-of-the-art NLP models. This is because it enables the model to “remember” all the words in the input and focus on specific words when formulating a response.



PES UNIVERSITY

Design Approach

The Attention mechanism has revolutionised the way we create NLP models and is currently a standard fixture in most state-of-the-art NLP models. This is because it enables the model to “remember” all the words in the input and focus on specific words when formulating a response



Design Constraints, Assumptions & Dependencies

Assumption:

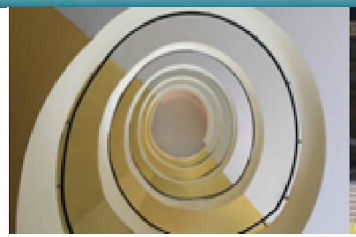
- Only certified radiologists will use the product.
- The images along with the reports received in the datasets are genuine and true.

H/W limitation: Storing of humongous amount of medical data.

Usage Limitation:

- The data cannot be made public, since it is of medical use.
- Restriction to unauthorised users, since it can be used in unethical manner to make money.

PES UNIVERSITY






Implementation Details

- Download and Prepare IU-Xray dataset: First, we web scraped the dataset from Open Access Biomedical Image Search Engine using basic Python modules and hen we split the dataset into 90:10 ratio for training and testing. There are a total of 7470 images in the dataset.
  Total Line of Code: about 72 lines

- Preprocess and Tokenization: We have used python file system to open the file, and then used basic string methods to preprocess the captions.
  For tokenization we have used Keras inbuilt tokenizer and text_to_sequence to add <unk> token and padding.
  Total Line of Code: about 33 lines

Implementation Details

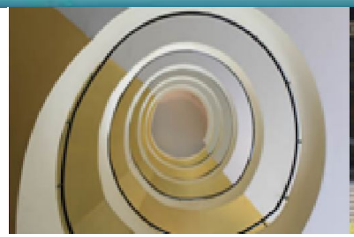
- Preprocess the image using InceptionV3:We have changed the input image size to 299px x 299px .
  Preprocess the images using the preprocess_input method from keras.application module of keras. to normalize the image, so that it contains pixels in the range of -1 to 1
  Total Line of code: 6 lines

- Initialize the V3 and pretrained weights: We have used pre-trained InceptionV3 model from keras.application module of keras. The weights are initialized as per ImageNet specification. We discarded the last layer as it is used for classification. We have also encoded the images and passed it through this model to get feature vectors.
  Total Line of code: 4 lines

Implementation Details

- Encoder and Decode: The CNN layer can be thought as enoder. For both encoder and decoder we hae use tf.keras.model, which provides with inbuilt functionalities.We have use Bahadnau Attention algorithm in our RNN layer for working with longer captions.
  Total Line of code: 45 lines

- Training: For training we have used tf.GradientTape() and tf.expand_dims which are the functions provided by tensorflow.
  Total Line of code: 32 lines






Project Results & Discussion

Reference Paper	Cosine Similarity	BLEU	ROUGE	METEOR
Frequency Model	NA	0.442	0.187	0.176
Nearest Neighbour	NA	0.281	0.209	0.125
CNN-RNN(Vinyals et al, 2015)	NA	0.316	0.267	0.159
LRCN(Donahue et al, 2015)	NA	0.369	0.278	0.155
Soft Att(Xu et al, 2015)	NA	0.399	0.323	0.167
ATT-RK(You et al, 2016	NA	0.369	0.323	0.171
CNN-LSTM(Sonit Singh, 2019)	NA	0.374	0.307	0.163
Co-Attention(Jing et al, 2018)	NA	0.369	0.447	0.217
Our Model	**0.325**	**0.452**	**P-0.308 R-0.328 F-0.267**	**0.201**

Project Results & Discussion

The outcome of our model seems encouraging, but there are certain issues which need to be addressed before deploying the automated report generation into the picture. The reasons being-

1) The current RNN networks are not good enough for long radiology reports, because they are unable to save the long range dependencies.
2) There is a need for separate evaluation metrics for radiology reports. BLEU score works on the basis of overlap of words, which can calculate wrong accuracy. For captions *"There is no evidence of pneumothorax, pleural effusion, and cardiopulmonary abnormality"* and *"There is cardiopulmonary abnormality and pneumothorax. No evidence of pleural effusion"* the BLEU score can come out to be 100% which is wrong.

Lessons Learnt

Attention Mechanism:

We learnt that classical image captioning model(Plain Encoder-Decoder architecture) generates the next word of the caption, this word is usually describing only a part of the image. It is unable to capture the essence of the entire input image. Using the whole representation of the image to condition the generation of each word cannot efficiently produce different words for different parts of the image.This is exactly where an **Attention mechanism** is helpful.

Transfer Learning :

Because of **shortage of data** for our project we took help of Transfer Learning to use pre build deep learning network which would compensate for our lack of data.Using a **"pre-trained model"** often speeds up the process of training the model on a new task, and can also result in a more accurate and effective model overall.

Lessons Learnt

Some of the challenges we faced during the evaluation of the model were that there was not a evaluation metric present which could measure the accuracy our generated caption medically.

All the accuracy score which are present used the context of English language to check similarity between actual X-ray report and our generated caption.

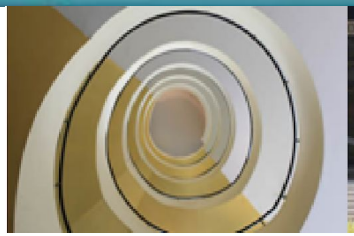
So researching on the correct accuracy score for medical purpose could be a whole project in itself. So we just used what all scores we found were more useful in our context.

Planned Effort Vs Actual Effort

The timelines for execution of the project:

Literature Survey	11th Jan
Data Collection/Preprocessing/ Exploratory Data Analysis	12th Jan-1st Feb
Model Building	2nd Feb-22nd Feb
Prototype and Fine Tuning	23rd Feb-21st Mar
UI Building , and Result Analysis	22nd Mar-11th Apr
Final Report	12th Apr-25th Apr

PES UNIVERSITY



Thank You