X-RAY IMAGE CAPTIONING

Group 1

Swati Sundar, Taylor Stevenson and Srivatsan Chakravarti

Northwestern University, MSDS 458 - Artificial Intelligence & Deep Learning

June 5th, 2022

Abstract

There has been a growing pool of research on machine learning in the medical field in recent years. Used in combination, techniques such as Natural Language Processing (NLP) and Deep Learning can assist medical personnel in making quicker and more accurate decisions. We will be performing image captioning tasks using these techniques. Image captioning is the automatic generation of descriptions of medical images, such as Magnetic Resonance Images (MRIs). This research effort aims to generate reports of chest X-Rays initially using pre-trained models to perform description extraction and then using that info to generate caption using LSTMs and GRUs. The dataset being used is publicly available from Indiana University and consists of chest X-ray images and reports. The goal is to predict the physician's notes of the information associated with the image. We conducted a series of 8 experiments and compared the performance of each experiment in this paper. The structure of our Deep Neural Network consists of three main models :- A CNN: used to extract the image features, An Encoder: The extracted image features are then passed to a Transformer, A Decoder: This model takes the encoder output and the text data. We will experiment with global attention techniques, pretrained ImageNet models, pretrained word embeddings and global flow and context flow techniques. The experimental results reveal that, to a certain extent, a deep neural network can be used to solve text captioning problems in the real world.

Introduction

Physicians and radiologists need to examine medical and radiological images and write their findings as medical reports. Any opportunity to assist doctors with describing findings will reduce medical errors and reduce the time spent and overall cost of care, benefiting the hospital. However, the process of writing these reports for a large number of cases on a given day can take a lot of the medical professional’s time. Automatic generation of image captions, describing the content of image, recognition, and localization of specific

diseases and organs is possible now with advances in computer vision and machine translation.

Generating a description of an image is called image captioning. This problem is an image captioning task. The objective of the exercise is to predict the impression part of the medical report. We are using chest X-rays, and the goal is to predict the impressions on the medical report attached to the images. Image captioning requires recognizing the essential objects, attributes, and relationships in an image. It also needs to generate syntactically and semantically correct sentences. Deep learning-based techniques can handle the complexities and challenges of image captioning. We will use pre-trained models to get information from input images and feed it into LSTM or similar network architecture to generate the caption.

The Chest X-ray dataset from Indiana University has two sets of files, one contains an X-ray image of patients, and the other includes radiology reports of that particular patient in XML format ("Indiana University - Chest X-Rays (PNG Images)" n.d.). The report could be associated with more than one image. The XML fields are image_id, the caption of an image, indication of patient, findings, and impression. There are 7471 X-ray images and 3955 reports in the dataset. The impression is the target feature.

Literature review

Image captioning has been utilized in both language and fashion and is a crucial problem to solve as it has demonstrated itself applicable across very different fields of interest. For example, a group of students at various universities across China was able to bridge a connection between clothing images and human semantic understanding (Li et al. 2021). Their methods consisted of a CNN for attribute detection and encoding LSTM model for decoding the features to the language description. In another example, researchers from Southwest University in Chongqing, China, generated Chinese image captions of pictures from a large-scale artificially labeled dataset proposed in the 2017 AI Challenge competition (Pan et al. 2021). Their encoder of choice was the Inceptionv4 model, and the decoder used was I-GRU. Both these examples share a similar methodology pattern, including an encoder-decoder framework,

the BLEU metric, and the same decoder being an LSTM and/or GRU model. As both groups of researchers have demonstrated success with these methods, we will also follow a similar methodology pattern.

Semantic Segmentation is a key tool used in computer vision tasks in which it assigns semantic labels to each pixel in an image. It has a broad application in fields like automated driving, augmented reality, and medical imaging. In 2020, researchers proposed a design called Chained Context Aggregation (CAM). The intent with CAM was to be used for image segmentation, however in the GF/CF experiment in our paper, it will be used for image feature extraction. The main components of CAM that will be used in our research today are called Global Flow and Context Flow. The use of these components will be explained in further detail in the Methods section. (Tang et al. 2021)

Methods

The methods for the state of the work on our topic are based on the encoder-decoder models. We decided to conduct the experiment individually as below:

1. Classic encoder decoder Methodology – Srivatsan Chakravarti
2. Attention-based encoder decoder methodology and EDA – Swati Sundar
3. Global flow and context flow – Taylor Stevenson

We started our experiment with data extraction, EDA, and text preprocessing. Subsequently, we downloaded the IU chest X-ray dataset from Open-I. The dataset from Indiana University has two sets of files, one contains an X-ray image of patients, and the other includes radiology reports of that particular patient in XML format ("Indiana University - Chest X-Rays (PNG Images)" n.d.). There are 7471 X-ray images and 3955 reports in the dataset. A maximum of 5 images are associated with each report and few reports did not have any linked images. As shown in fig.1 below the image distribution is not uniform and 81% of the report has 2 images associated with them.

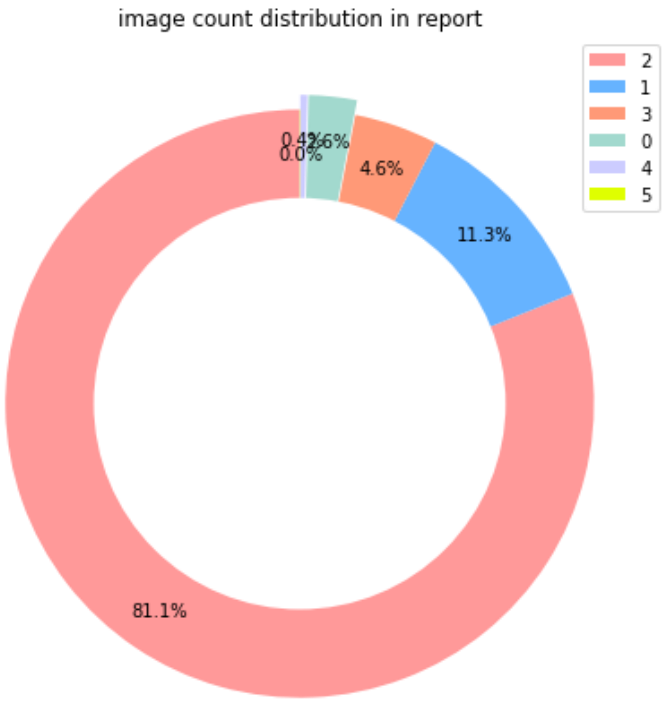


Fig.1 Image distribution in radiology report

We extracted all the information from xml report i.e., comparison, indication, findings, and impression part of the report and the corresponding heights, width, and image name of the concerned report to a dataframe with xml report file name. We created new datapoints with new images and same info for image count greater than 2 per report. We also performed text preprocessing like removing all numerical values like “1.” And “2.” , removing words like XXXX, removing full stop, removing unwanted spaces, expanded the contracted words like 'll to will and converted the final text to lowercase. As a part of data cleaning, we looked at the missing values on our extracted and preprocessed dataframe and dropped the null values for image 1 and impression column. we used the same image file in image 1 for missing values as in image 2. We also looked at the height and width distribution of images which gave us an idea on the shape of images. The height distribution was scattered with most values in size 420 and 512 while the width distribution was unique with 512 as shown in fig 2.

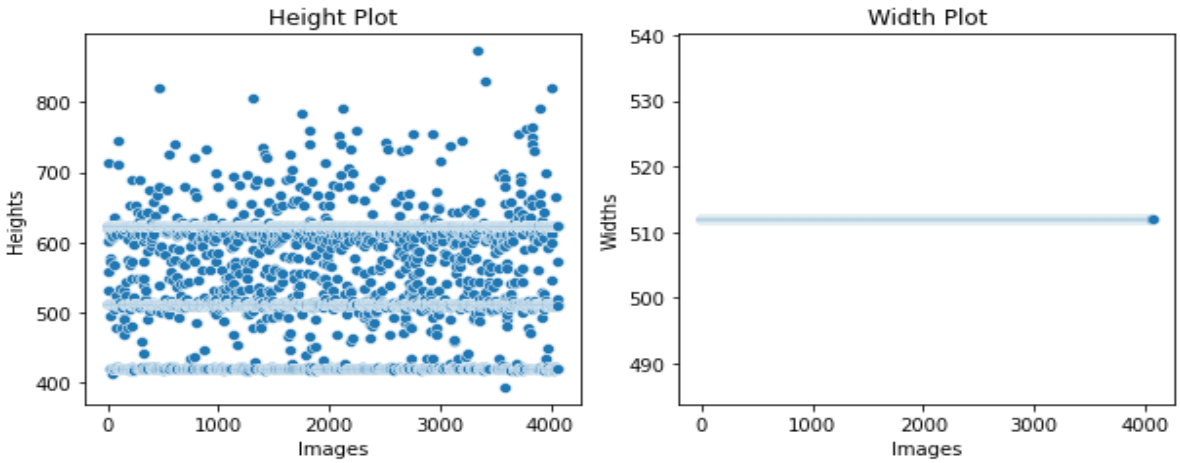


Fig.2 Image height and width distribution

We visualized the images and associated captions from our final dataset as shown in fig. 3. We also used wordcloud library to visualize the most common words present in the impression for radiology report as shown in fig. 4.

Fig.3 Image Caption in radiology report

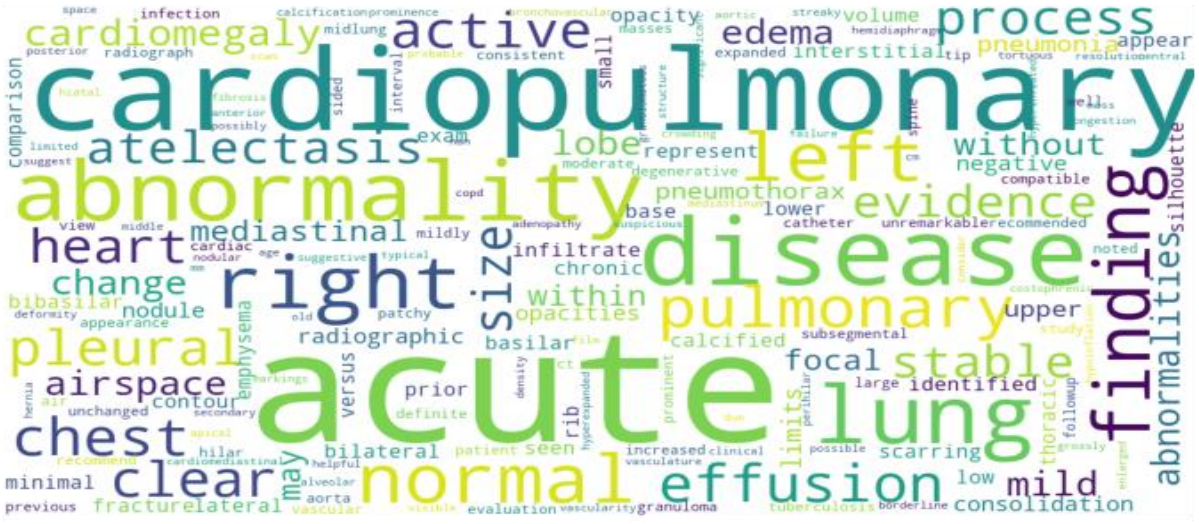


Fig.4 wordcloud

Then, we split the final preprocessed dataset into a train, test and validation set of 80/10/10 respectively with stratified sampling as the proportions in the target variable i.e., impression was very different with 60% of dataset with no disease label. Furthermore, we did up-sampling and down-sampling to reduce overfitting on this dataset.

For the Simple Encoder Decoder method, we conducted three experiments using different pre-trained models that allows global image features to be extracted from the hidden activations of CNN and then fed them into an LSTM to generate a sequence of words.

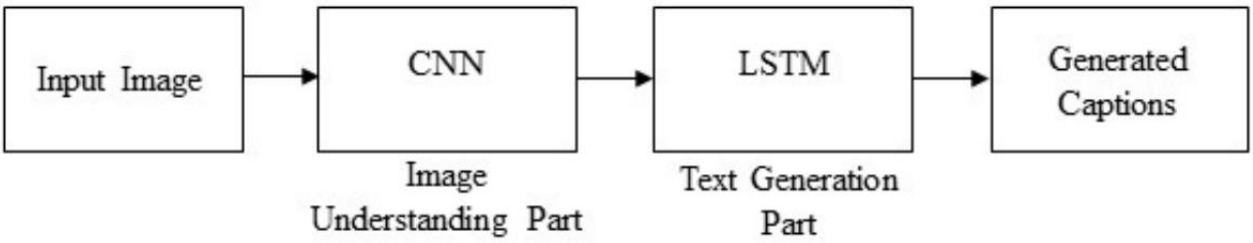


Fig. 5 Conceptual view of classic encoder decoder model

The method uses CNN for image representations and an LSTM for generating image captions. This special CNN uses a novel method for batch normalization and the output of the last hidden layer of CNN is used as an input to the LSTM decoder (Fig. 5). This LSTM can keep track of the objects that already have been described using text. In generating image captions, image information is included to the initial state of an LSTM. The next words are generated based on the current time step and the previous hidden state. This process continues until it gets the end token of the sentence.

Classic encoder decoder Methodology

The encoder part of the model will take the two images, image_1 and image_2 column in the dataframe and convert the images into features that can be provided to the decoder. In experiment 1, the encoder is a CheXNET model, which is a Denset121 layered model which is trained on millions of chest x-ray images for the classification of 14 diseases (Rajpurkar et al. 2017). The weight of that model is loaded, and the images are passed through that model. The top layer will be ignored. The CheXNet model 'trainable' parameter is set to false. The padded tokenized captions will be passed through an embedding layer, here we are using pretrained Glove vectors (300 dimensions) as the initial weights for the layer. This will be set as trainable and is passed through LSTM where the input to LSTM is taken from the output of the Image_dense layer. These are then added through an output dense layer where the number of outputs is the vocabulary size with SoftMax applied. 'Adam' optimizer with 'Sparse Categorical' is the loss function used for training the model.

We have used similar pre-trained models and loaded the weights of these models trained on ImageNet. For experiment 2 we have used ResNet50 for the encoding and is passed through a LSTM layer like experiment 1. For experiment 3 we have used pre-trained weights of VGG19 model.

Attention-based Encoder Decoder Model

A total of 4 experiments were conducted to test the attention-based encoder decoder network.

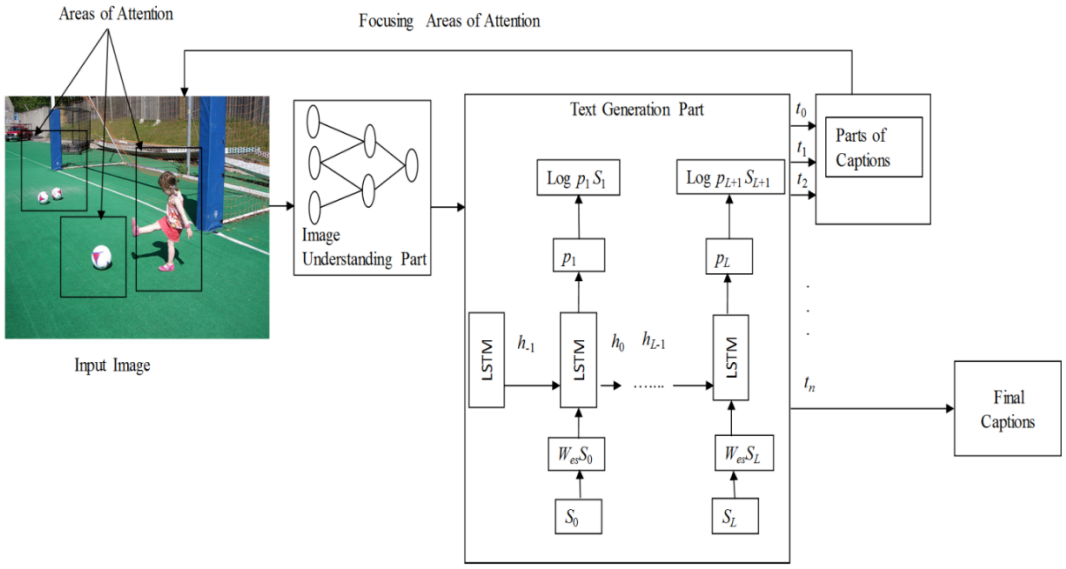


Fig. 6 Attention based image caption method

We experimented with global (Bahdanau) attention-based encoder and decoder model in experiment 1. The idea of a global attentional model is to consider all the hidden states of the encoder h_s when deriving the context vector c_t . Bahdanau attention or additive attention takes concatenation of forward and backward source hidden state. We used pretrained CheXNET model as an encoder and RNN (e.g., GRU) as a decoder with pretrained Glove Corpus of 300 dimensions. CheXNET is a Denset121 layered model pre-trained on millions of chest x-ray images to classify 14 diseases. During training, we used batch normalization, weight regularization and dropout techniques to regularize the model. The model gave an accuracy of 87.6% on the train set while 80% of accuracy on test and validation set, which shows signs of overfitting.

For experiment 2, We used global (Luong) attention based encoder decoder model with same architecture as in experiment 1. Luong's multiplicative style gives us local attention in addition to global attention as shown in the equation 1 below. The model almost gave same

accuracy as in experiment 1 of 79% for test and validation set, however train accuracy dropped to 73% showing signs of overfitting. Removing regularization like dropout, batch normalization layer and weight regularization introduced in experiment 1 might have helped in improving model's underfitting.

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases}$$

For experiment 3, We used inception v3 pretrained ImageNet model as encoder with same architecture as in experiment 1. The model almost gave the same accuracy as in experiment 2 of 79% for test and validation set, however train accuracy increased by 2% i.e., 75%. The model did show some signs of underfitting but not as in experiment 1. However, the Bleu evaluation score dropped, and the prediction was not as accurate as in experiment 1 & 2.

For experiment 4, We used spacy pretrained embedding matrix with same architecture as in experiment 1.Spacy is large pretrained corpus on English word. The model gave an accuracy of 81% on the training set and 80% on test and validation set. The issue with model's overfitting and underfitting was resolved, however bleu score and prediction accuracy dropped as compared to Glove embedding.

We also used a different test set of X-ray images with diseases like the Chest X-ray Images (Pneumonia) from the NIH for generating image captions to benchmark our model's performance("Box," n.d.). However, the model generated a caption as "No Acute cardiopulmonary abnormality" for images with disease label as shown in fig. 7.

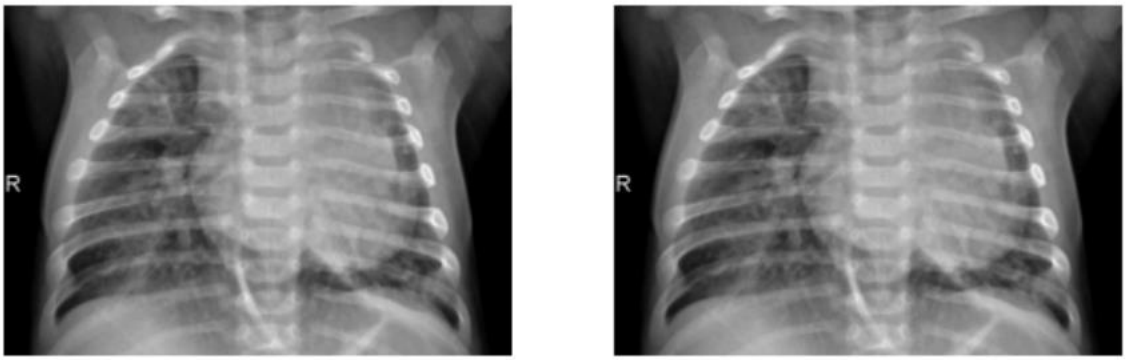


Fig.7 Predicted Caption on NIH Chest X-ray Image.

Global and Contextual Flow Model

This model was adopted from the research of Attention-Guided Chained Context Aggregation for Semantic Segmentation, as mentioned in the Lit Review. Fig. 7 below, provides an outline of this model's configuration:

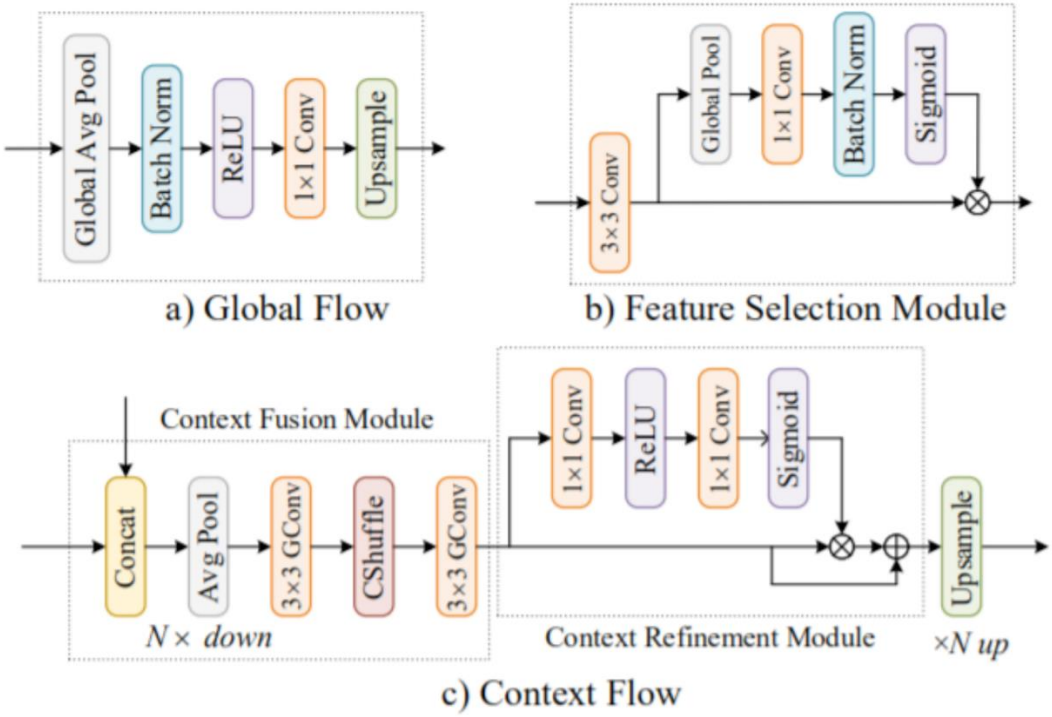


Fig. 7 Global context flow method

This model used a similar encoder/decoder as the Attention Model, with some modifications. The job of the encoder is to extract features from the output of Image_Encoder layer from the CheXNET model. Output from this model will then be passed through the Global Flow (GF) with the goal of extracting global features of each image. (Chempolil 2021) Again, both the output from CheXNET and the output from Global Flow will be passed through the Context Flow (CF), where local features will be extracted from each image. Ultimately, the output from the GF and the CF will be summed, reshaped, and normalized before being sent to the decoder. (Chempolil 2021)

Global Flow: Information from the Image_Encoder layer will pass through the global average pooling layer. Batch Normalization is applied, the activation function used is ReLU, and then the image is upsampled and shaped to match the size of the input. Conv1D was used to extract global image representations. (Tang et al. 2021)

Context Flow: This portion of the model will take both the outputs from Image_Encoder and GF and concatenate on the last axis. Global average pooling is applied, and this reduces the size of the feature map.

Result

Here are the results of various methods used for the predictions and evaluate to produce captions that predict the correct sentences. We would be using BLEU (Bi-lingual evaluation understudy) scores as one of the methods to evaluate the performance of these models. Individual text segments are compared with a set of reference texts and scores are computed for each of them. In estimating the overall quality of the generated text, the computed scores are averaged. However, syntactical correctness is not considered here. The performance of the BLEU metric is varied depending on the number of reference translations and the size of the generated text. BLEU is popular because it is a pioneer in automatic evaluation of machine translated text and has a reasonable correlation with human judgements of quality.

Classic encoder decoder method results:

For the classic encoder methods, we found the CheXNET model provided the best performance loss/accuracy scores and BLEU score metrics. We have used greedy search which takes our list of potential outputs and the probability distribution already calculated — and chooses the option with the highest probability based on the very next word or token. This can cause the method to get stuck on certain word or sequence and repetitively assigning these sets of words the highest probability. Table 1 below is a summary of the loss/accuracy scores of the three experiments, with CheXNET providing the best BLEU score results.

Experiments	**Val Accuracy**	**Val Loss**	**Train Accuracy**	**Train Loss**	**BLEU (1gram)**	**BLEU (2gram)**	**BLEU (3gram)**	**BLEU (4gram)**
CheXNET	0.4903	0.221	0.6561	0.2570	0.312769	0.3107	0.335934	0.3685
ResNet50	0.4734	0.2843	0.8688	0.0900	0.2364	0.2449	0.2969	0.3592
VGG19	0.4556	0.3076	0.8681	0.0899	0.2643	0.2478	0.3057	0.3416

Table 1- Classic encoder decoder results

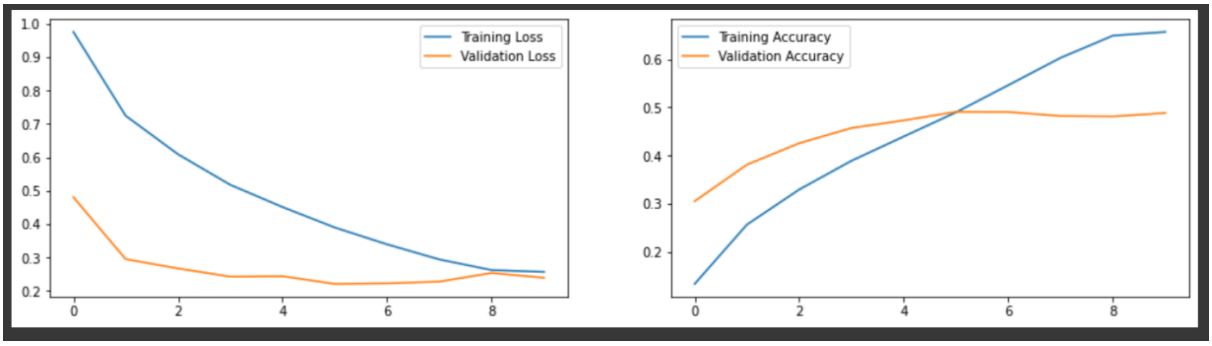


Fig. 8 Training accuracy/loss curve for Chexnet classic encoder/decoder

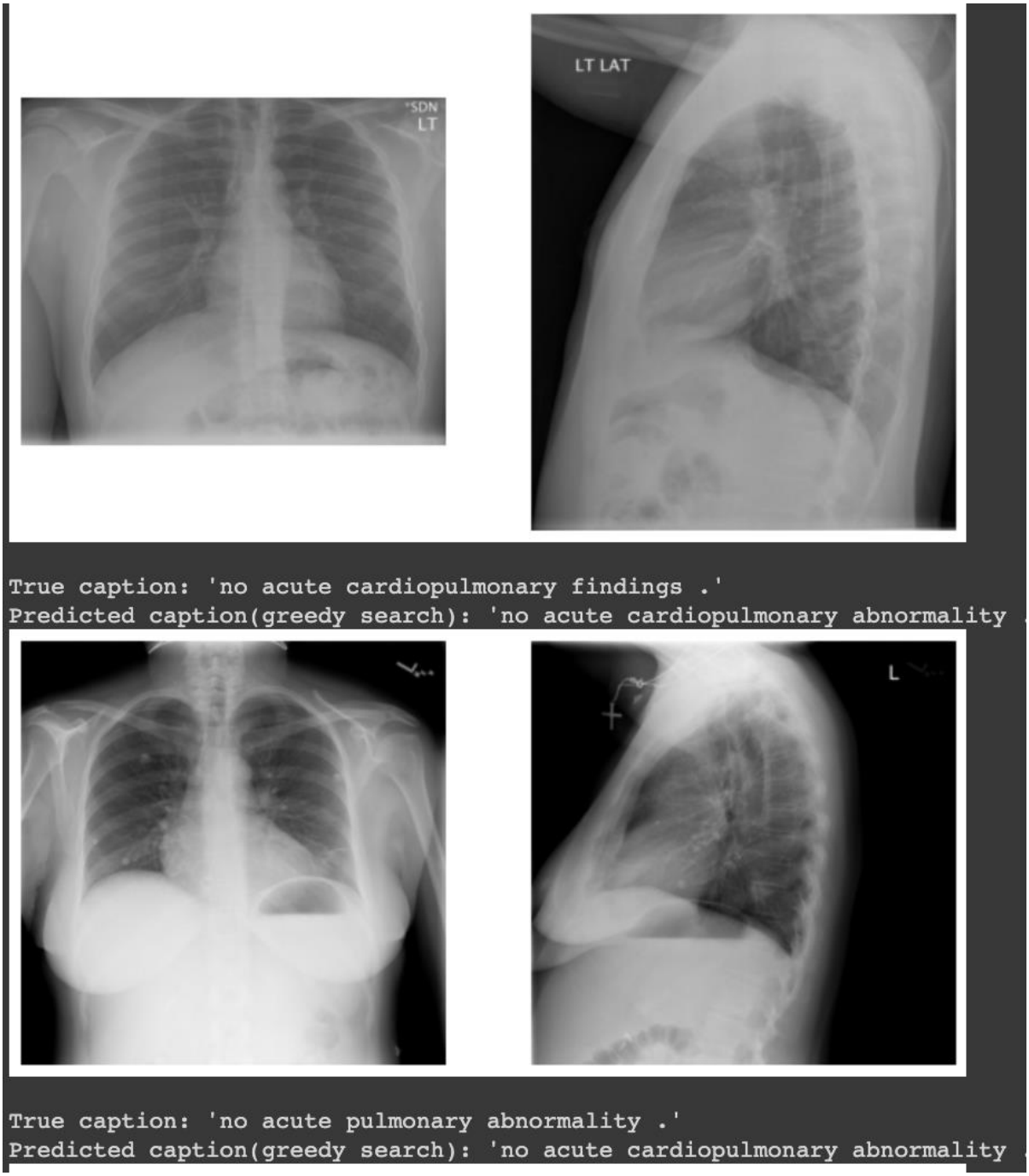


Fig. 9 – Predictions using simple Chexnet encoder/decoder method

Attention based encoder decoder method results:

The results in Table 2 below describe the training and test accuracy results for various attention-based techniques used along with encoder/decoder model. Taking BLEU scores together with test accuracy, Bahdanau and Spacy embedding did well compared with other attention models.

Model	Train-accuracy	val-accuracy	Test-Accuracy	Train-loss	val-loss	Test-loss	Bleu-1gram	Bleu-2gram	Bleu-3gram	Bleu-4gram
Bahdanau Attention - Encoder Decoder	0.87673	0.79892	0.8008	1.88359	1.99901	2.00602	0.3127	0.3165	0.3498	0.3867
Luong's Attention - Encoder Decoder	0.7333	0.7938	0.7917	2.2437	1.9892	2.0103	0.3255	0.3165	0.3429	0.3765
Inceptionv3 Encoder Decoder	0.75657	0.79325	0.78487	2.34998	2.10032	2.1364	0.2957	0.2953	0.3271	0.3631
Spacy Embedding - Attention Encoder decoder	0.81485	0.80117	0.80067	2.04853	1.980693	1.99334	0.3126	0.3165	0.3498	0.3867

Table 2 Attention based encoder decoder results

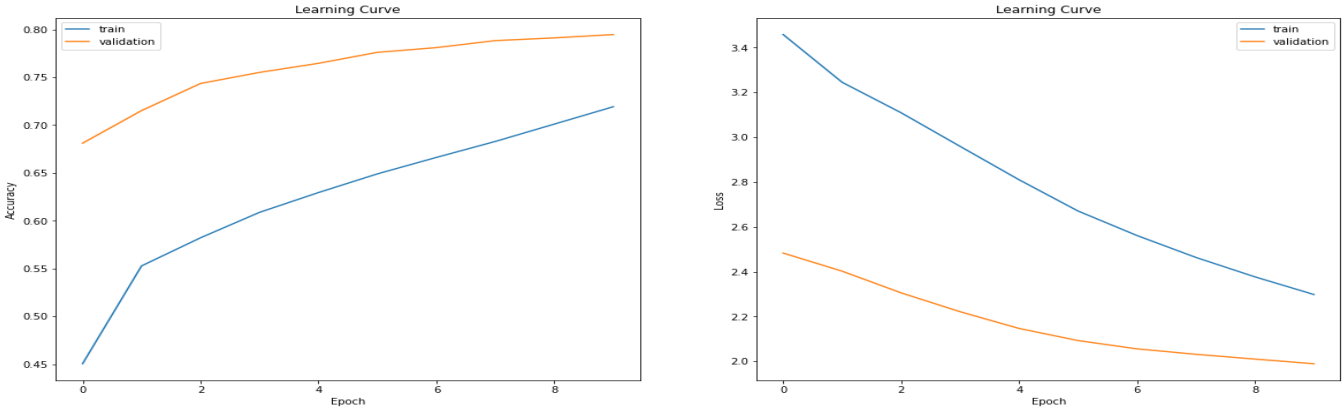
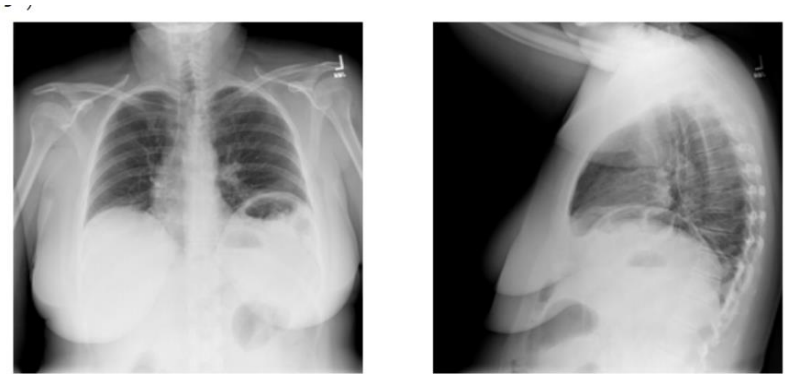


Fig. 10 Learning Curve - Luong's attention-based encoder/decoder



```
True caption: 'low lung volumes with streaky bibasilar opacities subsegmental atelectasis over infiltrate .'
Predicted caption(greedy search): 'low lung volumes with minimal left basilar atelectasis .'
Predicted caption(beam search = 1): 'low lung volumes with minimal left basilar atelectasis .'
```

Fig. 11 Predicted caption from Luong's attention-based encoder/decoder

Global and Contextual method results:

The results in Table 3 below describe the training and test accuracy results for various Global and context methods used along with encoder/decoder model.

Experiment	Train Loss	Train Accuracy	Val Loss	Val Accuracy	Run Time	Bleu-1gram	Bleu-2gram	Bleu-3gram	Bleu-4gram
Global & Context Flow	0.14	0.95	0.89	0.77	20 min 30 sec	0.212536	0.238073	0.290892	0.349789

Table 3 Global Contextual flow based encoder decoder results

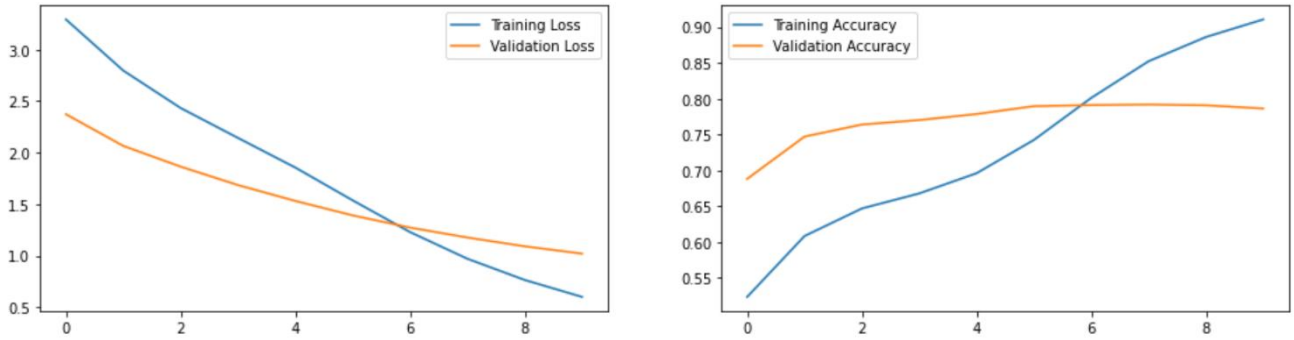


Fig. 12 Learning Curve - Global Contextual flow based encoder decoder

Conclusion

In this paper, we conduct a series of experiments implementing a Encoder Decoder Transformer Neural Network to achieve the text generation of the IU Chest X-Ray database. The results of Global Attention and Global Contextual Flow Encoder Decoder model are promising. However, most of the prediction resulted with "no acute cardiopulmonary abnormality". A balanced dataset with more variability, specifically X-rays with disease captions, would have been helpful for the models to understand better and can help reduce bias towards the “no disease category.

For future work, we can use word embeddings pretrained on radiology reports or medical vocabulary like BioBert pretrained model. We could also try using more balanced dataset on radiology report to reduce bias towards on disease category.

References

"Indiana University - Chest X-Rays (PNG Images)." n.d. Academic Torrents. Accessed May 9, 2022. https://academictorrents.com/details/5a3a439df24931f410fac269b87b050203d9467d.

Pan, Yongbin, Lidan Wang, Shukai Duan, Xiuling Gan, and Liangyi Hong. 2021. "Chinese Image Caption of Inceptionv4 and Double-Layer GRUs Based on Attention Mechanism." Journal of Physics: Conference Series 1861 (1): 012044. https://doi.org/10.1088/1742-6596/1861/1/012044.

Li, Xianrui, Zhiling Ye, Zhao Zhang, and Mingbo Zhao. "Clothes Image Caption Generation with Attribute Detection and Visual Attention Model." Pattern Recognition Letters 141 (2021): 68–74. https://doi.org/10.1016/j.patrec.2020.12.001.

Tang, Quan, Fagui Liu, Tong Zhang, Jun Jiang, and Yu Zhang. 2021. "Attention-Guided Chained Context Aggregation for Semantic Segmentation." Image and Vision Computing 115 (November): 104309. https://doi.org/10.1016/j.imavis.2021.104309.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, et al. 2017. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning CheXNet." https://arxiv.org/pdf/1711.05225.pdf.

"Box." n.d. Nihcc.app.box.com. https://nihcc.app.box.com/v/ChestXray-NIHCC.

Chempolil, Ashish Thomas. 2021. "Medical Image Captioning on Chest X-Rays." Medium. February 9, 2021. https://towardsdatascience.com/medical-image-captioning-on-chest-x-rays-a43561a6871d.