Semantic Analysis of Chest X-Ray using an Attention-based CNN Technique

Rishabh Dhenkawat[1], Snehal Saini[2], Nagendra Pratap Singh[1]
[1]Department of CSE, National Institute of Technology, Hamirpur (H.P.),India
[2]Department of ECE, National Institute of Technology, Hamirpur (H.P.),India

Keywords: CNN, LSTM, Deep Learning, Additive Attention, Teacher Force.

Abstract: The world today is suffering from a huge pan-demic. COVID-19 has infected 106M people around the globe causing 2.33M deaths, as of February 9, 2021. To control the disease from spreading more and to provide accurate health care to existing patients, detection of COVID-19 at an early stage is important. As per the World Health Organization (WHO), diagnosing pneumonia is the most common way of detecting COVID-19. 172K deaths were reported in the USA between February 2020 and January 29, 2021, that was caused by pneumonia and COVID-19 together. In many situations, a chest X-ray is used to determine the type of pneumonia. We present a deep learning model to generate a report of a chest x-ray image using image captioning with an attention mechanism.

1 INTRODUCTION

Computer vision-based diagnosis provides an automatic classification and suggestions for reference to improves diagnosis's accuracy and efficiency. In the past few years, many deep learning and machine learning algorithms are used for the classification of medical images, SVM, K-nearest neighbors, random forest, and other techniques are included. They can be used in a variety of medical image processing applications.

Using old machine learning methods poses two major difficulties. First, the inaccurate results due to the limited processing of large input. Secondly, the use of manual feature extractions instead of learning valid features. Thus, deep learning methods are preferred for medical image processing.

learning's technology has a wide variety of applications in healthcare image processing, such as diagnosis and organ segmentation. The convolution neural network cnn has been used extensively in several pieces of research that include reading and interpreting ct images for medical applications. Deep learning is a representation learning technique that connects different layers and nonlinear components efficiently to obtain various representation levels.

Deep learning algorithms have two essential characteristics: local connectivity and shared weights (CNN). Deep learning is widely used in image analysis because of all these features, which make it much easier to handle complex data processing tasks. Convolution layer, pooling layers, and fully connected layers are the three layers that make up the CNN architecture. Convolution layers extract features from the previous layer, pooling layers minimize computational complexity, and completely connected layers, eventually, are used to extract features from the previous layer. A recurrent neural network (RNN) is used to process sequence data in order to recognize things. Since words in a sentence are semantically related, word generation uses previous word knowledge to predict the next word in the sentence. In RNN, the current output of a sequence is related to the previous output, enabling word relationships to be determined. It is used to model temporary sequences and their long-range dependencies because of the property of feedback connections.

In this paper, we propose a CNN-LSTM chest-x-ray image semantic analysis focused on an attention process to produce a description of the chest x-ray images. In the deep learning model, we used the idea of the attention to highlight the infection regions in the lungs. Two types of attention mechanisms in deep learning are local attention and global attention. In our pour model, we used Local Attention, also known as additive attention or Bahdanau Attention.

Dhenkawat, R., Saini, S. and Singh, N.
Semantic Analysis of Chest X-ray using an Attention-based CNN Technique.
DOI: 10.5220/0010567200003161
In *Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering (ICACSE 2021)*, pages 218-223
ISBN: 978-989-758-544-9

As a result, the model assists in the analysis and clarification of chest x-ray images, automatically supplying doctors with valuable knowledge about the input x-ray image. Two types of chest x-ray images available are frontal and lateral sides. Using these two types of images as data, our model generates a report for these chest x-ray images. To construct a deep learning model, we present a predictive model that uses both image and text processing. This paper uses chest X-ray images from Indiana University's large Chest X-Ray dataset to describe the model's architecture and detection efficiency.

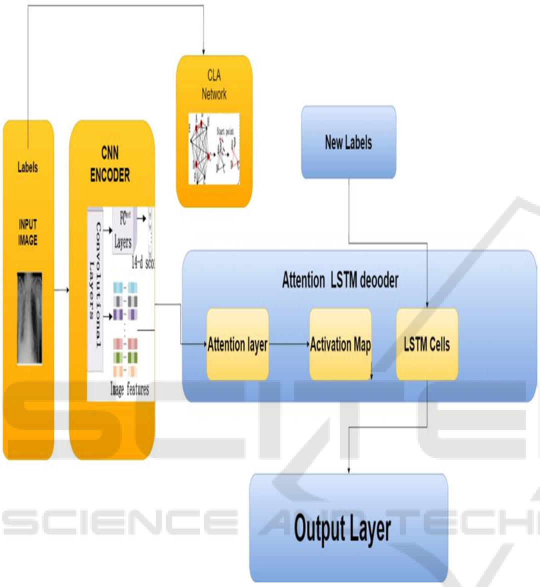


Figure 1: Model Flow

2 METHODOLOGY

2.1 Overview

We have used an encoder and a decoder architecture with an attention mechanism and compared it with encoder and decoder architecture without attention. Here Convolution Neural Network is taken as an encoder to extract visual features, this encoder will output image feature vectors. Resulted feature vectors will be taken as input to an additive attention-based LSTM decoder. LSTM decoder took image feature vector and sequence vec-tor to process reports. An image classification using InceptionV3 model over chest dataset is used with this classification model the weights were saved over the training and later used in Encoder feature extraction by using the saved weights to InceptionV3.

2.2 CNN Encoder

The Convolution Neural Network is popular in deep learning due to its ability to learn and represent image feature vectors. Many frameworks like VGG16, Resnet, Inception, and Densenet are trained on Imagenet Dataset containing 1.3 million natural images. Due to the difference between medical chest images and natural images, In-ceptionV3 is again trained on labeled chest x-ray images to improve transfer efficiency. The encoder is a single linear model which is fully connected. The input X-ray image is fed to InceptionV3 which extracts the features of two images are adds them. Then they are input to the FC layer and an output vector is obtained. The encoder's last hidden state is connected to the Decoder.

2.3 LSTM Decoder

Recurrent Neural Networks models the non-static behavior of sequences through connections between different units. LSTM is a type of RNN which have 3 added states as forget state, input state, and output gates. Hence the LSTM layer is present in the decoder which does language modeling up to word level. The first step receives encoded output from the encoder and the ¡start¿ vector. The input is passed to the LSTM layer with additive attention. The output vector is two vectors one is the predicted label and the other is the previous hidden state of the decoder, this feedback goes again to the decoder on each time step.

2.4 Attention Mechanism

Attentive neural networks are used in wide-ranging applications like summary formation, translation, photo captioning, etc. They act as tools to take account of hidden feature maps which make networks analyze important regions. It provides weights to each channel in the feature map. Attention is of 2 types:

Global Attention (Luong's Attention): Attention is placed on each and every source position.

Local Attention (Bahdanau Attention): Only some of the source positions receive attention.

In this work, we are using Local Attention(known as Bahdanau Attention) or additive attention which is placed only on a few source positions. As Global attention takes account of all sources side words for all target words, it becomes computationally very expensive and not efficient when translating long sentences. To address this problem, additive attention is employed, which focuses only on a small portion or subset of the encoder's hidden states per target word.

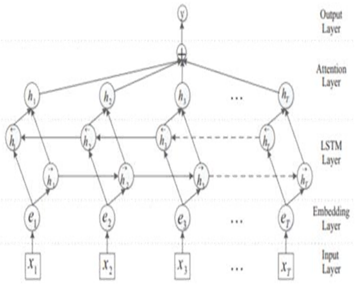


Figure 2: LSTM with attention

2.5 Model Architecture

The model proposed in this paper contains five components: Input layer input labels are given to the model and then summed with the image feature vectors. The embedding layer is then used to map each label to a low-dimensional vector. LSTM layer is used to get high-level features, from step these LSTM layers are repeated twice to understand features in more depth. A weight vector is provided by the Attention layer, and it also merges word-level features from each time step into a sentence-level feature vector, by multiplying the weight vector. Finally, the sentence-level feature vector in the output layer is finally used for relation classification.

3 EXPERIMENTS

3.1 Dataset

We have used Indiana University's vast chest X-rays dataset provided by the Open-i service of the National Library of Medicine. The dataset contains 7000 chest x-rays from various hospitals along with 3,900 associated radiology reports. Each report is associated with two different views of the chest, i.e., a frontal and a lateral view. The associated tags contain the basic findings from the x-ray images which are used to train the model so as to generate image captions later on.

3.2 Exploratory Data Analysis

Before jumping to the main code, we analyzed the dataset to visualize some of its important characteristics. For eg, by performing text analysis on the impression column target variable we got the bar plot of the most unique sentences for indication in the x-ray reports and the frequency of their occurrences.

By generating a word cloud we can see the most occurring words in the sentences present in x-ray reports. Some of these words are chest pain, shortness, breath, male, female, dyspnea, and indication. The word cloud is used to represent the words having the maximum word count in the impression column target variable.

Further, we visualize the word count distribution plot for the impression column target variable. This plot offers better insights to see the minimum and maximum word count. From the plot, we conclude that the minimum word count is 1, the maximum word count is 122 and the median word count is 5.0.

Further, we analyze the distribution of image count per patient using a bar plot and we see that the minimum image count is 1 and the maximum image count is 5.

Since two types of chest x-ray images are available to us which are frontal and lateral view. By selecting a sample data point we find out the total number of images present for that particular patient, its findings, and impressions. From here, we analyze that there are multiple images associated with every patient.

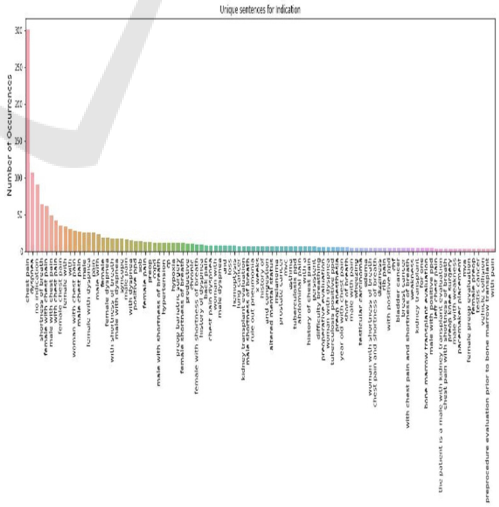


Figure 3: Bar Plot of unique sentences for indication

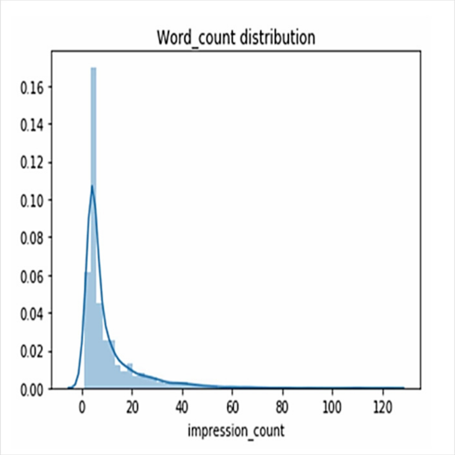


Figure 4: Word count distribution plot

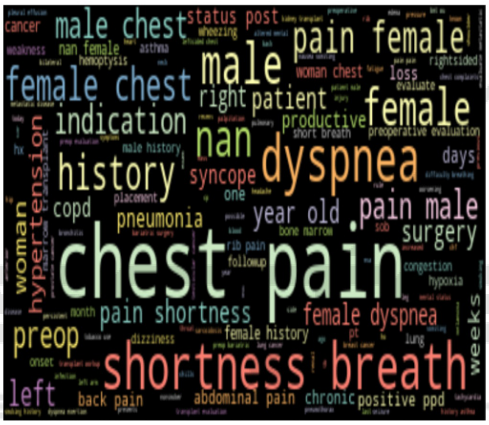
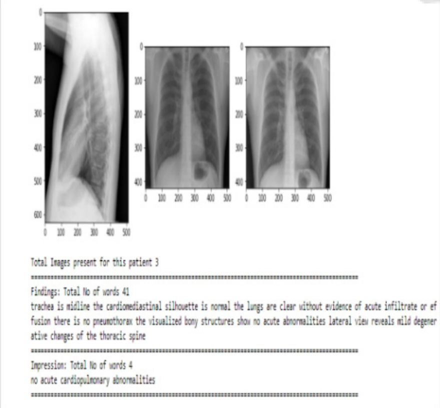


Figure 5: Word cloud for impression column



```
Total Images present for this patient 3
==================================================================================================
Findings: Total No of words 41
trachea is midline the cardiomediastinal silhouette is normal the lungs are clear without evidence of acute infiltrate or ef
fusion there is no pneumothorax the visualized bony structures show no acute abnormalities lateral view reveals mild degener
ative changes of the thoracic spine
==================================================================================================
Impression: Total No of words 4
no acute cardiopulmonary abnormalities
==================================================================================================
```

Figure 6: Results for a sample data point

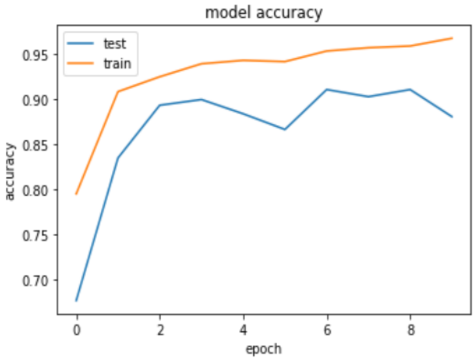


Figure 7: Model Accuracy for prepossessing

3.3 Pre-processing and Training

The transfer learning method is used for the image to feature vector conversion and text data tokenization is used for dataset preparation. InceptionV3 model trained on imagenet is used. A classifier to detect which type of disease the person is suffering from was made. Once classification was done weights of the trained model were saved in hd5 format.

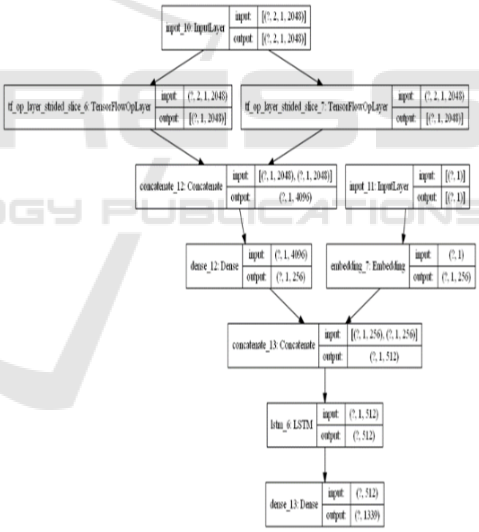


Figure 8: CNN-LSTM without attention architecture

3.4 Model without Attention Mechanism

3.4.1 Encoder Architecture

Single fully connected layer linear output is used. Before we pass to the FC layer, two image tensors were added and pass to FC layer. This layer outputs the shape of batch size and embedding dimension.

3.4.2 Decoder Architecture

It contains an embedding LSTM layer and dense layer which outputs shape (batch size, vocab size).

3.4.3 Model Training

In the training phase, the Teacher forcing is used. for training recurrent neural networks that use the output from a previous step as an input. In training, a “start” token is used to start the process and the generated word in the output sequence is used as input on the subsequent time step along with other input like an image or a source text. Until the end, the same recursive output as the input method is used till better results are generated.

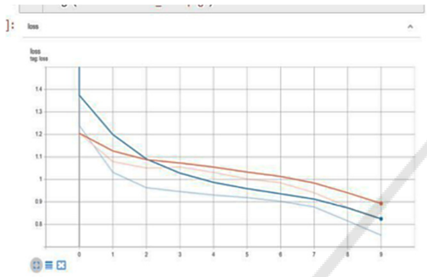


Figure 9: CNN-LSTM without attention loss

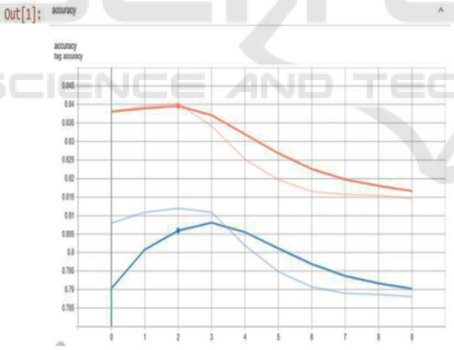
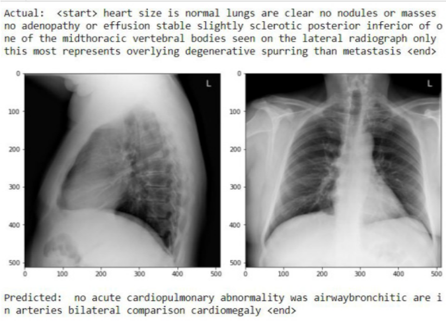


Figure 10: CNN-LSTM without attention accuracy



```
Actual:  <start> heart size is normal lungs are clear no nodules or masses
no adenopathy or effusion stable slightly sclerotic posterior inferior of o
ne of the midthoracic vertebral bodies seen on the lateral radiograph only
this most represents overlying degenerative spurring than metastasis <end>
```

```
Predicted:  no acute cardiopulmonary abnormality was airwaybronchitic are i
n arteries bilateral comparison cardiomegaly <end>
```

Figure 11: CNN-LSTM without attention resulted report

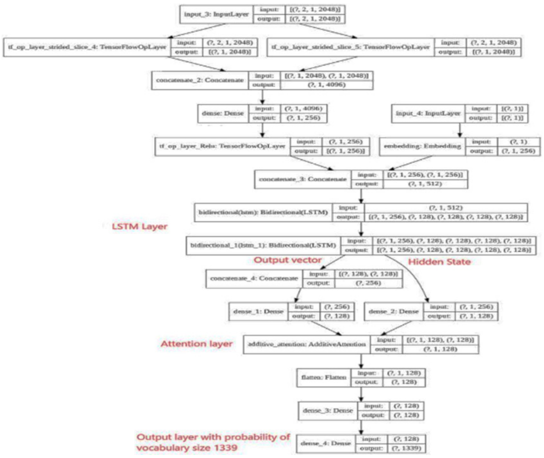
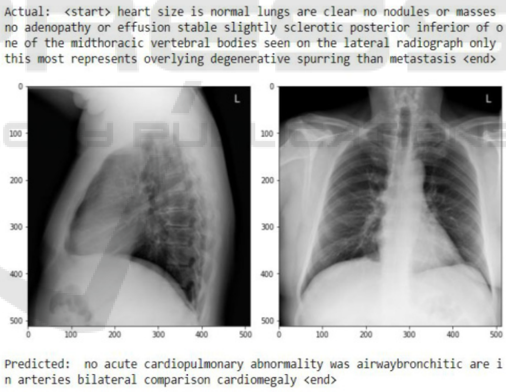


Figure 12: CNN-LSTM with attention architecture

3.5 Model with an Attention Mechanism

The encoder part is the same as the previous model architecture and summed image vector with a single fully connected layer. In the decoder part lstm with attention is used. Here additive attention (Local or Bahdanau attention) is used.



```
Actual:  <start> heart size is normal lungs are clear no nodules or masses
no adenopathy or effusion stable slightly sclerotic posterior inferior of o
ne of the midthoracic vertebral bodies seen on the lateral radiograph only
this most represents overlying degenerative spurring than metastasis <end>
```

```
Predicted:  no acute cardiopulmonary abnormality was airwaybronchitic are i
n arteries bilateral comparison cardiomegaly <end>
```

Figure 13: CNN-LSTM with additive attention resulted report

3.5.1 Model Evaluation

Beam search-based teacher forcing method is used to find the resulting sentence. Beam search is used here, as it chooses the most probable next step when the sequence is made. It uses all possible next steps and takes most likely k. Here k is user-specified and controls the number of searches.

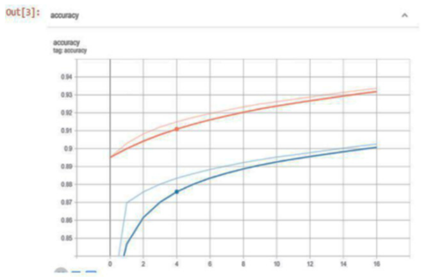


Figure 14: CNN-LSTM with additive attention accuracy

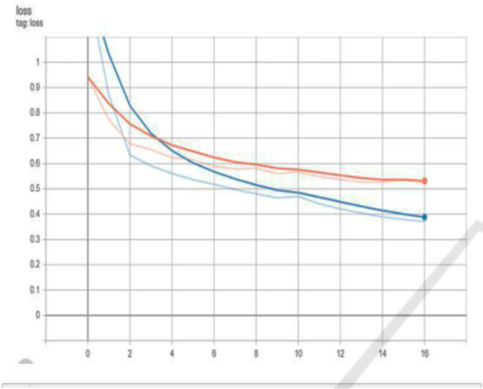


Figure 15: CNN-LSTM with additive attention loss

4 CONCLUSIONS

By comparing the results, Model Architecture of attention-Based Long Short-Term Memory Networks for Relation for the Classification worked well in classification tasks than without attention. Loss is converged to 0.3with an accuracy of 89 percent train and 92 percent validation from the result we can see there is a similarity between each predicted and actual output. Thus, by using the attention mechanism, along with conventional deep learning methods, we can improve the accuracy of the model.

ACKNOWLEDGEMENTS

This paper and the research behind it would not have been possible without the exceptional support of my supervisor, Dr. Nagendra sir. His enthusiasm, knowledge, and exacting attention to detail have been an inspiration and kept my work on track from our coding to the final draft of this paper. The magnanimity and proficiency of one and all have enhanced this study in innumerable ways and saved us from many errors.

REFERENCES

Bahdanau, D., Cho, K. H., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Paper presented at 3rd International Conference on Learning Representations, ICLR 2015, San Diego, United States.

CNN+CNN: Convolutional Decoders for Image Captioning. / Wang, Qingzhong; Chan, Antoni B

Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., and Cohen, W. W., “Review Networks for Caption Generation” 2016.