First Review Document

MEDICAL REPORT GENERATION ON CHEST X-RAY IMAGES

R.S. Rahul Sai
17BCE0136
9131038763
rs.rahulsai2017@vitstudent.ac.in

Prof. Sharmila Banu K (11989)
Associate Professor Grade 1
9942469321
sharmilabanu.k@vit.ac.in

B.Tech.

in

Computer Science and Engineering

School of Computer Science & Engineering

Abstract

Computer Aided Diagnosis (CAD) and medical imaging systems have evolved in the past decade to a point where they can partially mimic radiologists and doctors. These systems can learn and differentiate the features and abnormalities in medical images, and provide objective evidence with a higher diagnostic confidence and faster inference. In this project, we focus on generating detailed medical reports on chest X-ray images, which can be adapted later to work with other diagnostic tools such as ultrasounds and mammograms. The Indiana University dataset provides us with CXR images corresponding to various lung and heart ailments, along with well-defined reports and findings. The generation of medical reports mainly consists of two broad tasks. The first task is to treat the problem as a multi-label classification task to obtain the accurate tags for a particular image from the visual features. This is performed using Convolutional Neural Networks and a transfer learning framework called ChexNet, which is specialized for chest X-ray images. The second task is to generate the reports using these aforementioned tags, which requires the use of recurrent neural networks such as hierarchical LSTMs. To improve the quality of the sentences produced, a co-attention mechanism is also required, which makes use of the spatial information from the convolutional layers and the generated words in order to find localized regions from which the abnormalities are found.

Introduction

The use of deep learning in image captioning has been a popular use case over the past few years. It is a challenging problem as it requires the machine to generate textual description from the contents of an image, similar to how a human brain would describe an image. But consider the same scenario for medical images, with the example of Chest X-ray images. For a normal human eye, chest X-ray images are just images with the skeletal and muscular features of the lungs defined in black and white. But highly trained radiologists who have studied and diagnosed various respiratory and cardiovascular abnormalities, can mark multiple areas of the images and can write down clear reports for potential abnormalities. However, to read a chest X-ray image properly, it is important that the radiologist has a thorough knowledge of the human thorax, and how various respiratory diseases might affect them. This comes with multiple years of

experience by analyzing chest x-ray images with a fixed pattern of evaluation, and evolves over time depending on the history of cases a particular radiologist may handle. But, even for highly trained and experienced radiologists, writing reports is highly time-consuming especially in regions with higher population density. Looking at the other side of the spectrum, radiologists or pathologists in rural areas, with inferior imaging equipment face a similar issue. They either cannot get objective evidence to diagnose anything properly or lack the knowledge of the respiratory or cardiovascular abnormalities. Misdiagnosis of symptoms and medical errors are the third-leading cause of deaths across the world, leading to more than 3 million deaths. This project focuses on generating detailed medical reports using Chest X-ray images, which can facilitate the diagnosis of various respiratory and cardiovascular diseases. But to achieve this objective, we need to overcome several challenges.

A medical report can usually consist of various parts, which are usually heterogenous in nature. There may be images, abbreviations and complicated terminology in these reports. To avoid this issue, we will focus on three sections from a medical report i.e., findings – a large paragraph which contains keywords and parameters, indication – the doctor's advice for the patient and impression – a sentence finalizing the results for a report. To achieve this, we propose a multi-task architecture which works by predicting the keywords/tags (findings) by treating it as a multi-label image classification task and generate longer descriptions by using a text generating mechanism such as hierarchical LSTMs. Hierarchical LSTMs **[1]** are specialized recurrent neural networks which are often used for text generation from images and video frames. It is built to consider both high-level language features from the training text and low-level visual features obtained from the processed image.

The dataset to be used in this project is the Indiana University Chest X-Ray (CXR) Image dataset **[2]**. It is a high resolution CXR dataset with multiple views i.e., frontal, side and posterior views. There are 7,470 images accompanying 3,955 well written reports encoded in XML. These XML reports have references to the CXR images, the findings, impressions and the indication from the CXR images. These CXR images are obtained from patients diagnosed with tuberculosis, pneumonia and various heart ailments.

Frequent Tags	Total	Overlap
Normal	2696	0
Opacity	840	665
Cardiomegaly	655	490
Calcinosis	551	441
Hypoinflation	537	357
Calcified Granuloma	508	300

Table 1 : Most Frequent Tags occurring in the dataset

Objectives

1. Preparing an NLP Pipeline for the original findings, impressions and indication
   a. Removing Contractions, Punctuations and Numbers
   b. Tokenization
   c. Representing the reports in word embeddings
2. Obtain the image features using a Convolutional Neural Network and the Transfer Learning framework i.e., ChexNet (acts as the encoder here).
3. Generate the text using the labels/tags obtained from the encoder using Hierarchical LSTMs which act as the decoder in our Sequence-to-Sequence Model.
4. Substituting Attention mechanisms to improve the encoder-decoder approach and compare the results.
5. Improving the results by utilizing newer NLP techniques such as BERT Embeddings **[3]** and Transformers **[4]** and generate better reports.

Problem Statement

To generate detailed medical reports with three heterogenous sections i.e., Findings, Impressions and Indications from IU Chest X-Ray images, highlighting the context of the particular disease i.e., location, severity and affected organs, using a combination of CNN, hierarchical LSTMs and co-attention mechanism.

Literature Review

In **[5]**, the authors discuss the viability of a cascade model for medical image captioning where they cascade CNNs and RNNs over multiple steps. The first step is to train the CNN with the images and predicting single object labels, and the RNN to describe their context from the text. The second step carries over the weights from the previous step and introduces a mean pooling layer in order to derive the image/text context labels from the image/text contexts of the previous step. The type of RNNs used in this framework are often used sequence generation types i.e., LSTMs and GRUs, which use the input image's context vectors (in the form of CNN embeddings) in order to learn the annotation sequence.

The concept of using semantic attention is discussed in **[6]** where the authors first discuss the top-down paradigm (start from the high level features of the image and come up with words) and bottom-up paradigm which starts with words which describe various features of the image and combines them to form a coherent sentence using language models. Both paradigms suffer from their own weaknesses such as lack of attention to fine details in the top-down approach and the lack of end-to-end procedures from the individual features to sentences in the bottom-up approach. They suggest an idea that visual attention plays a major role of offering feedback which can help combine both the top-down and bottom-up information. Visual attention can be defined briefly as the mechanism in our visual cortex which tends to look at the low-level and semantic details of the image which it considers more important rather than the whole image. Their approach involves detection of semantic attributes using the bottom-up approach, which they call candidates, and then they employ a top-down approach in order to select which candidates should require more attention in order to yield better results. Their framework outperforms competing methods across various evaluation metrics such as BLEU and Meteor.

A similar concept of caption generation is discussed in **[7][8]** where the authors discuss the concept of scene understanding with the help of visual attention. They propose two techniques under a common framework, one being a soft and deterministic attention mechanism and another being a hard stochastic attention mechanism, which can be trained by maximizing a convergence function. A CNN extracts a 14x14 feature map, which is then processed by a RNN with visual attention over the image which provides a context vector. This vector is processed by a word LSTM which generates a word-by-word caption by utilizing a greedy search mechanism.

System Architecture

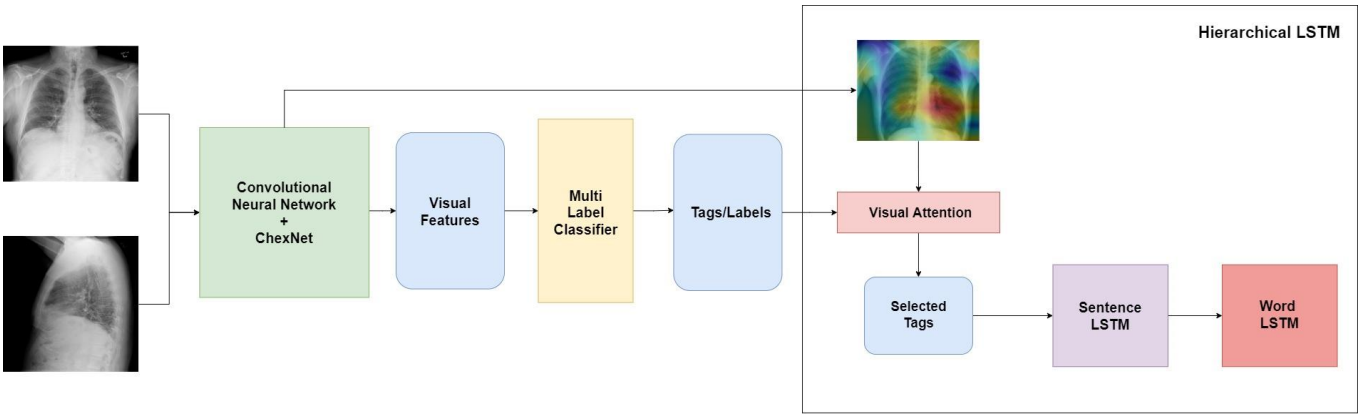

Figure 1 : Architecture for the medical report generating system

The architecture of the medical report generation can be divided into three distinct parts:

1. **NLP Pipeline** – The findings, impressions and the indications obtained from the reports have to be properly cleaned to be used in the model. This involves the following steps:
    a. Converting all characters into lowercase
    b. Removing contractions from the text e.g., won't – will not, can't – cannot.
    c. Removing punctuation from text with the exception of full stop, as the findings from the reports may contain more than one sentences.
    d. Removing all numbers and redacted data from the text.
    e. Removing smaller words and adverbs with the exception of "no" as it adds significant value. e.g., adverbs such as "there", "then".
    f. Tokenization and addition of identifier tokens such as "_start" and "_end" tokens which are necessary in the text generation process.

Examples –

Findings before the cleaning process:

No pleural effusion or pneumothorax. No acute bone abnormality.

Findings after the cleaning process:

_start no pleural effusion no pneumothorax. no acute bone abnormality. ***_end***

2. **Convolutional Neural Networks –** In order to extract the visual features of the image, we make use of convolutional neural networks. The CNN acts as the encoder in our model, and provides us a feature vector with the visual features of the image, which can be used to predict the tags for that particular X-ray. This is done by considering the problem as a multi-label classification task where the output layer provides you with the probability of a set number of tags. These predicted tags help us massively in text generation which is discussed in the next section. However, the size of our dataset (7,470 images) is not enough to train a CNN properly. To remove this bottleneck, we have to use a transfer learning framework. Most transfer learning frameworks such as VGG16 or InceptionV3 are trained over generic image datasets which doesn't serve our purpose.

   Fortunately, ChexNet is a convolutional neural network especially trained on Chest X-ray images. It was trained over 1,12,120 images and contains 121 layers where the input is a chest X-ray image, and the output is the probability of 14 different diseases along with a localized heatmap which highlights the visual features of the chest x-ray image. However, we do not need to classify the image into one of those 14 categories, so we can remove the final classification layer. From a image of dimensions (224,224,3), we get a feature vector with a length of 1,024. We have two images associated with a report, so we concatenate the two feature vectors to get a feature vector with length 2,048. This final feature vector will be passed along with the report to the decoder which is discussed in the next section.

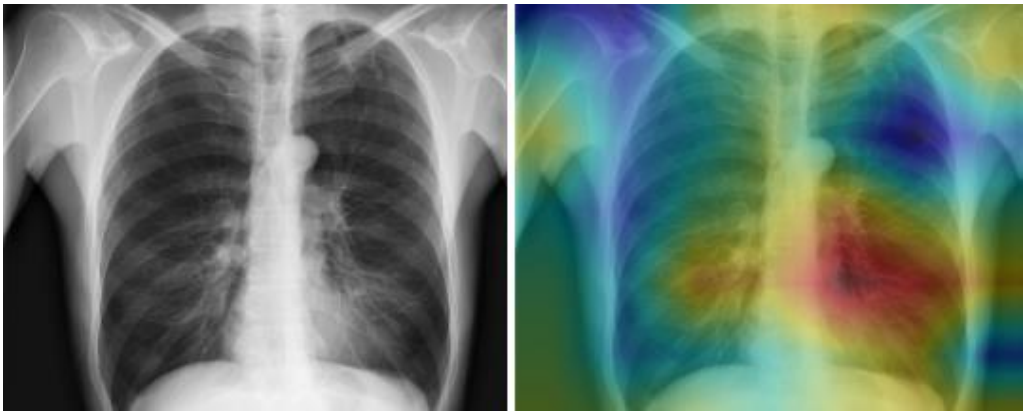

Figure 2: Chest X-ray image from the dataset before and after passing through the CNN

3. **Hierarchical LSTMs and Attention Mechanism –** Hierarchical LSTMs [1] are specialized recurrent neural networks which are often used for text generation from images and video frames. It is built to consider both high-level language features from the training text and low-level visual features obtained from the processed image. Note that the keywords/tags obtained from the image are generated by the fully connected layer, which results in a loss of spatial information. To improve these results, an additional mechanism known as Co-Attention is added. Co-Attention mechanism uses the spatial information from the visual features of the convolutional layers and the semantic features obtained from the tags of the specific image (which are generated by the fully connected layer).

   When the visual features and the tags arrive at the decoder, the high-level spatial information provided by the localized heatmap help us to focus on the tags which are visually highlighted more, and yield better results. This new context vector with the embeddings of the selected tags is passed on to a sentence LSTM. A Sentence LSTM will generate multiple sentences as suggestions to the provided words, using a technique known as beam search which predicts the probability distribution across the given vocabulary and returns the words which have the maximum probability to be subsequent words in the sentence. Beam search selects multiple alternatives for an input sequence, based on conditional probability and a parameter known as beam width. When the sentence vector is successfully produced, it is passed on as a context vector to the Word LSTM. Word LSTM employs a greedy search mechanism, which selects a single candidate which is suitable for the input sequence in a time step. This improves the quality of the final sentence generated.

Conclusion

The automated generation of medical reports on chest X-ray images is highly viable, using a multi-step approach which employs the best techniques offered by deep learning and NLP. Chest X-rays are one of the most popular diagnostic tools and this project can improve the speed and quality of diagnosis The CNN predicts the tags from the visual features and retains the spatial information in order to provide better context. The hierarchical LSTMs can decode the vector provided by the encoding layer in order to generate legible medical reports, which can be compared and evaluated using metrics such as BLEU and Meteor Scores.

References

[1] L. Gao, X. Li, J. Song, and H. T. Shen, “Hierarchical LSTMs with adaptive attention for visual captioning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, 2019.

[2] D. Demner-Fushman *et al.*, “Preparing a collection of radiology examinations for distribution and retrieval,” *J. Am. Med. Informatics Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “{BERT:} Pre-training of Deep Bidirectional Transformers for Language Understanding,” *CoRR*, vol. abs/1810.0, 2018.

[4] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[5] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, “Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2497–2506.

[6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image Captioning with Semantic Attention,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4651–4659.

[7] K. Xu *et al.*, “Show, attend and tell: neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, 2015, pp. 2048–2057.

[8] B. Jing, P. Xie, and E. Xing, “On the Automatic Generation of Medical Imaging Reports,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2577–2586.