Medical Image Captioning Using Attention Mechanisms

Devipriya Raju

Abstract

Modern image captioning techniques, as noted in (Lei et al.), mostly concentrate on enhancing visual features; little emphasis has been made to leveraging the intrinsic capabilities of language to improve captioning performance. In this essay, we demonstrate how creating high-quality image captions requires enhanced image processing, consideration of the language processing and the syntactic paradigm of sentences. We examine various approaches to captioning medical images that incorporate attention mechanisms, according to the traditional encoder-decoder structure. Although the idea of image captioning is not new, the medical field has not yet used it to its full potential. This paper is a naive attempt to further that area of study. We implement greedy and beam search with attention mechanism with regards to the captioning process.The base code for this project is derived from (Vinitha, 2021).The other materials referred for this project are (Vidya, 2020) and (Ashish, 2022)

1 Introduction

Today's diagnostic workflows require the use of medical imaging. Because it is affordable and convenient, X-Ray continues to be one of the most often utilised visualization techniques among the variety of imaging modalities currently available in hospitals all over the world. For identifying and keeping track of a variety of lung conditions, including pneumonia, analysing and interpreting X-ray pictures is very important.The creation of a free-text description based on the results of clinical radiography is now a useful tool in clinical practice. Radiologists are overburdened by the requirement to submit their observations in writing, a laborious and time-consuming activity requiring in-depth domain-specific expertise, while also having to analyze about 100 X-Rays every day. This typical manual annotation overload can result in a number of issues, including missed findings, inconsistent quantification, and a patient's hospital stay being prolonged, which raises the cost of the therapy. The ability of radiologists to establish accurate diagnoses should be cited as one of the biggest issues.As mentioned in (Selivanov et al., 2022) there is a greater requirement for a strong image captioning framework in the COVID-19 era. As a result, many healthcare institutions contract out the duty of medical picture analysis. Deep learning-based automated chest X-ray medical report production can help and speed up the clinicians' procedure for making diagnoses. Automating this task has the potential to simplify clinical operations and raise the standardization and quality of care.

2 Literature Survey:

2.1 Image Processing:

The process of converting an image into a digital format and carrying out specific procedures to extract some usable information from it is known as image processing. When implementing specific specified signal processing techniques, the image processing system typically interprets all images as 2D signals.

2.2 CNN Algorithms:

As mentioned in (Khan et al., 2019) unique class of neural networks called Deep Convolutional Neural Network (CNN) has excelled in competitions involving computer vision and image processing. Image Classification and are two of CNN's intriguing application areas.Speech recognition, object detection, segmentation, video processing, and natural language processing. Deep CNN uses many feature extraction phases to automatically learn representations from the data, which contributes to its strong learning capabilities. The research on CNNs has advanced because to the availability of a lot

of data and advancements in hardware technology, and interesting deep CNN architectures have recently been described. A number of innovative concepts for improving CNNs have been investigated, including the use of various activation and loss schemes.

2.3 InceptionV3:

InceptionV3 (Szegedy et al., 2015) is an architecture that was basically formed with a series of convolution layers that includes the design principles like factorizing convolutions with large filter size, factorization into small convolutions, spatial factorization into asymmetric convolutions and utilization of Auxiliary classifiers.

2.4 DenseNext121:

Densenet121 model is a model of 5 layered dense blocks built from 121 layers. In this model, each layer connects to other layer through a feed forward fashion.The feature-maps of all layers before it are utilised as inputs for each layer, and its own feature-maps are used as inputs into all levels after it. DenseNets offer a variety of compelling benefits, including the elimination of the vanishing-gradient issue, improved feature propagation, promoted feature reuse, and a significant decrease in parameter requirements.This model has been described in detail in (Huang et al., 2016)

3 Related Work

In (Sehgal et al., 2020), the CNN and RNN are adopted. The image is initially delivered as input through CNN like InceptionV3 to determine the image's context and the preprocessed model makes use of transfer learning. As the output of the CNN model, the words are given back as a set. The system is then communicated with via Natural Language Processing (NLP). RNN is ultimately guided with the aid of the Flickr8k text data set. The inspected objects are consequently delivered to the RNN after undergoing specific planned operations, and the RNN then generates some pertinent and important caption.An essential goal of this research is to introduce the deep learning approach for producing captions using neural networks, as described in Recognition and Detection of Objects in Generation of Image Caption System (Kumar et al., 2019). The proposed way is to produce captions as in perception and the identification of the items using deep learning. This includes Handcrafted Feature Extraction, Deep Feature Extraction Using CNN Method, Face Region Detection and Normalization Pre-Processing, and Image Feature Concatenation, Selection, and Detection. In (You and Zhao, 2019) images and word vector pairs are supplied to the GRU during training, and at each stage, the GRU uses the same parameters. The probability distribution of the subsequent word, which is described as a fixed vector in the word dictionary, is produced by GRU. Making the maximum likelihood obtain the maximum value is the aim of training. In this case, the loss function is be the negative sum of each anticipated description of the training images.

In order to enhance performance of the image identification, researchers have recently concentrated on adding background data to the model. The authors take the same course in this work(**?**). The research proposes to simultaneously learn features and context information utilising the new Global-Local Attention method, in contrast to earlier efforts that learn the context information individually. They contend that because there is a correlation between local and global information, the accuracy of the network can be increased by simultaneously learning the attention utilising both. The topic of this work is emotion detection, and it is based on the idea that, in addition to the expression on a person's face, additional context clues, such as a gesture or a position, can also reveal the emotion of that person. Also I used certain methodologies to address the issues of imbalance in the data set as explained in (Kotsiantis et al., 2005). The process of under sampling is a a non-heuristic technique called random under-sampling seeks to balance class distribution by arbitrarily removing examples from the dominant class.

4 Methodology:

Though there were related work as listed above, they were concentrating on general images that consist of an object or a person or an animal. The image captioning process on medical-related images were very less and difficult process as the difference between 2 images is very less. In other words every images of a particular type of scan have a,lot of similarity. This project focuses on various attempts of image captioning on a series of radiology images from (rad).

Initially, we chose InceptionV3 model with a minimal Exploratory Data Analysis and data modification. The initial choice of Inception V3 was due

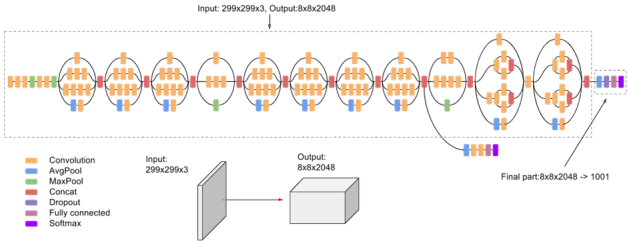

Figure 1: InceptionV3 Architecture

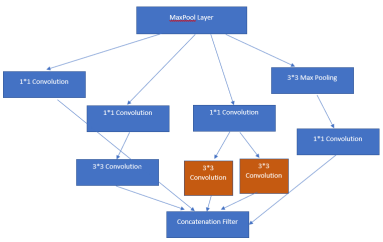

Figure 2: Modified InceptionV3 Architecture

to the fact that has got lower error margin comparatively.Hence, we trained the InceptionV3 model with pre trained weights-Imagenet and Attention mechanism. Then we extended our experiment with simple Densenet121 model. These both experiments provided a very poor performance. During the Exploratory Data analysis, I found that there is a huge imbalance of the data set. Hence, to address that we used down sampling and up sampling of certain columns according to it's data they represented. After many efficient data analysis, we chose to series of experiments to compare the results and chose the best model. Initially,we again chose InceptionV3 with and without pre trained weights were used on the images for feature extraction process. Then I chose to run Densenet121 with this cleaned dataset. This CNN architecture was initially trained by xray images only and hence the idea of training this model without pretrained weighst was discarded.

The next set of experiment was done with a slightly modified InceptionV3 model built from scratch. This model is an naive attempt to reduce the number of parameters trained in an efficient way, so that the training time can be reduced dramatically. Though this model has issues with other kinds of image classification that comes with certain shapes, these issues were minor. Since, we have already tried the with a combination of InceptionV3 models, this was assumed to give an extra mileage for our performance.

A modified version of InceptionV3 where I have replaced a 5*5 convolution from Inception block A, 1*1 convolution from Inception block B and a 3*3 convolution block from inception block C are replaced by pairs of 3*3-3*3,1*7-7*1 and 3*1-1*3 respectively in order to reduce the parameters.

A traditional Encoder-Decoder architecture was used for this image captioning project.These CNN architectures were entangled in an Image Encoding block. This block extracts the features from both

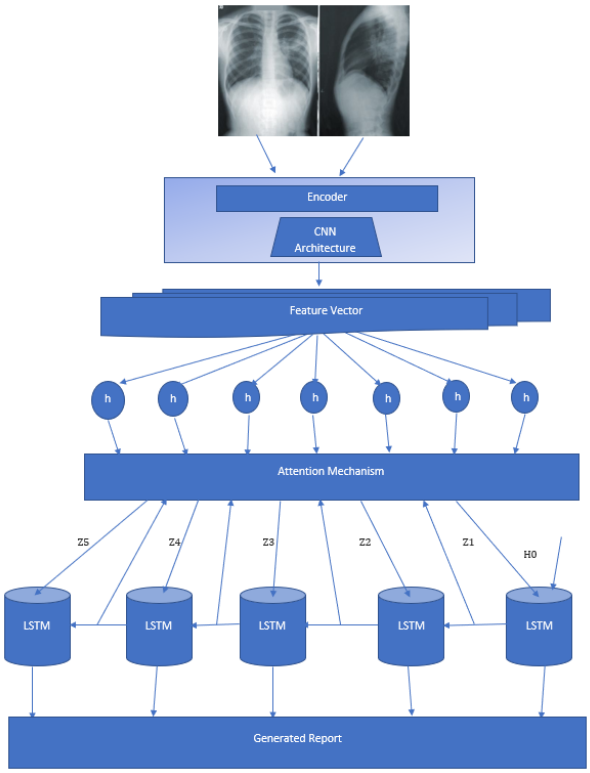

Figure 3: Encoder-Decoder with Attention Mechanism'

images of each report and then passes it as an input to the Decoder. This enclosed architecture for image captioning purpose was proposed by (Bappy et al., 2019) and it was handy to use in this type of project as it enabled us to experiment with a series of models with changes in few lines of code.

A global attention mechanism which captures the whole context of the image and the texts from the LSTM decoder have been used in this project. The attention weights, context vector and the attention vector have been calculated according the given formulas in the Figure 4. The decoder works on these inputs to generate the captions.

Beam and Greedy search have been used to predict the captions of the X-ray images. Greedy search selects the text based on the calculated conditional probability of each word from the vocabu-

$$\alpha_{ts} = \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad \text{[Attention weights]} \quad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \quad \text{[Context vector]} \quad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t]) \quad \text{[Attention vector]} \quad (3)$$

Figure 4: 'Attention Mechanism'

lary(1).

$$y_t = argmax_(y \epsilon V) P(y|y_1, y_2 .... y_{t-1}, x) \quad (1)$$

Beam Search selects the highest probability of sequences that results from greedy search(2)(3). Since the data set is huge, it takes a long time to calculate Beam search and hence, is is not considered it for bulk prediction.

$$log(ab) = log(a) + log(b) \quad (2)$$

$$logP(y_1, y_t ... y_T|x) = \Sigma_{t=1}^{T} logP(y_t|y_1, y_2, ... y_{t-1}, x) \quad (3)$$

5 Results:

The basic InceptionV3 and Densenet121 models trained with data set on which we conducted only basic EDA resulted very poor captions.To our surprise, the model trained without pre trained weights produced substantial empirical results. The model caught the abnormalities that even the original captions missed. Though the similarity between generated and actual captions were less, the model still did a decent job on the captions with a decent BLEU score. But the ones that had lesser BLEU score, the model did a pretty job on predicting the captions.

The modified InceptionV3 when trained only on these radiology images, didn't have a overall good BLEU score. But did a great job on identification of the abnormalities. In many cases, the predictions and the actual captions meant the same but were not exact word to word. Also, the model found the abnormalities which were present in the x-rays, but were not mentioned in the actual data set. This means that even for the model can absorb some of the features from the abnormality in the images and could able to identify that on the other images. For example, consider the following figure. Though the actual and prediction captions are not the same, the predicted caption is also not wrong. In other words, we can say that the model still have a lot to improve.

Method	$BLEU_1$
InceptionV3 with Pretrained Weights	0.00071
InceptionV3 without Pretrained Weights	0.193
Densenet121	0.246299
Simple InceptionV3	8.061e-232
Simple Densenet121	8.07e-23

Table 1: BLEU Score for Greedy Search predictions

Method	$BLEU_2$
InceptionV3 with Pretrained Weights .	5.3425e-157
InceptionV3 without Pretrained Weights	0.08
Densenet121	0.14569

Table 2: BLEU 2 Score for Greedy Search predictions

Also the caption "no acute cardiopulmonary abnormality" is widely available in the data set. Though, we addressed the imbalanced data set issue, still this problem exists and hence this appears as mostly predicted one.

Conclusion

We have tried different models and mechanisms for this Xray image captioning. The modified version of InceptionV3 exhibits great results comparatively. Though this was not shown out rightly from the BLEU Score, this can be proven from the human verification perspective. As an improvement of this project, we can implement GPT2 model in the decoder along with BERT embedding system.

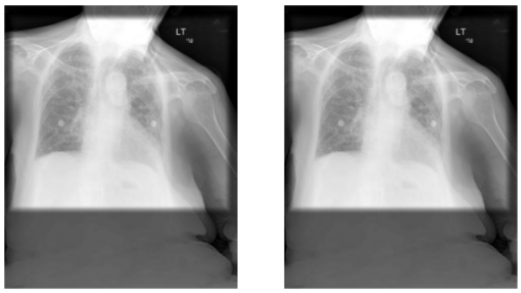

Figure 5: True caption: 'changes of chronic interstitial lung disease with ill defined patchy left apical and right basilar airspace disease . pa and lateral chest radiograph may be of benefit clinically feasible .' Predicted caption(greedy search): 'no acute cardiopulmonary abnormality . technically limited exam . incidental note of large cervical spine osteophytes .'

References

Ashish. 2022. Medical report generation. https://github.com/ashishthomaschempolil/Medical-Image-Captioning-on-Chest-X-rays/blob/main/EDA_Medical_Report.ipynb.

Jawadul H. Bappy, Cody Simons, Lakshmanan Nataraj, B. S. Manjunath, and Amit K. Roy-Chowdhury. 2019. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely connected convolutional networks. *CoRR*, abs/1608.06993.

Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. 2019. A survey of the recent architectures of deep convolutional neural networks. *CoRR*, abs/1901.06032.

Sotiris Kotsiantis, D. Kanellopoulos, and P. Pintelas. 2005. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36.

N. Komal Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj. 2019. Detection and recognition of objects in image caption generator system: A deep learning approach. In *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pages 107–109.

Ke Lei, Pei Wenjie, Li Ruiyu, Shen Xiaoyong, and Yu-Wing Tai, editors. *Reflective Decoding Network for Image Captioning*.

Smriti Sehgal, Jyoti Sharma, and Natasha Chaudhary. 2020. Generating image captions based on deep learning and natural language processing. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 165–169.

Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V. Dylov. 2022. Medical image captioning via generative pretrained transformers.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Analytics Vidya. 2020. A hands-on tutorial to learn attention mechanism for image caption generation in python. https://github.com/Vinithavn/Medical-report-generation/blob/master/EDA.ipynb.

Vinitha. 2021. Medical report generation. https://github.com/Vinithavn/Medical-report-generation/blob/master/EDA.ipynb.

Fucheng You and Yangze Zhao. 2019. Attention image caption with densenet. *Journal of Physics: Conference Series*, 1302:032048.