Automatic Report Generation for Medical Images

Muhammad Kamran[1] Mohib Ullah[2] Ali Shariq Imran[2] Muhammad Sajjad[2]
[1]Islamia College University Peshawar, Pakistan.
[2]Norwegian University of Science and Technology, Norway.

Abstract
In this work, we propose an encoder-decoder-based automatic report generation system capable of generating radiology reports for chest x-rays. We tested five backbone Convolutional Neural Networks, namely VGG16, InceptionV3, Resnet50, MobileNet and NasNet mobile, to extract visual features and used Long Short-Term Long Memory (LSTM) to extract the text features from the reports. Both features are concatenated and given to a deep network for report generation. We performed experiments on publicly available Indiana University's NLMCXR dataset. We evaluated our system against different backbones and evaluated accuracy and BLEU score. The result showed that our method achieved reliable and convincing results.

Keywords
X-ray, Convolutional Neural Network, Long Short-Term Long Memory, encoder-decoder, visual features, medical reports.

1 INTRODUCTION

Medical images like radiological, X-ray, CT, and MRI are popular in the medical field due to their usefulness in diagnosis and prognosis [1]. Looking at and interpreting these images is an arduous, tedious, and time-consuming task for medical experts. In the medical field, time is of prime importance because saving patients' lives is the doctors' primary objective.

Automatic Report Generation is a task similar to image captioning. Primarily, visual features are extracted from images using computer vision techniques and the text features from corresponding captions for the images using natural language processing (NLP) and combining those visual and text features to generate a caption for the corresponding image [2]. It is widely used to assist visually challenged people with hearing problems. The idea of generating a report from medical images is inspired by recent work in caption generation for natural images using multi-modalities, i.e., natural language data and images data done by Andrej et al. [3]. They used Flickr8, Flickr30k and MSCOCO datasets for experimentation. Their model was based on a combination of Convolutional Neural Network (CNN) [4] for visual features extraction and bidirectional Recurrent Neural Networks for extracting features from text data. Yuan Xue et.al [5] proposed a novel generative model that automatically generates a complete radiology report. Their proposed model combines the Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) in a recurrent way. Their model cannot only generate high-level conclusive impressions but can also generate a detailed description of findings sentence by sentence to support the conclusion. The multi-modal model combines the encoder with an image and one generated sentence to build attention input so that it can guide the generation of the following sentence. In this way, the model maintains the coherence among the generated sentences. Omar et al. [6] proposed a conditioned transformer-based method to generate radiology reports and claim that they are the first to condition the pre-trained transformer on visual and semantic features to generate medical reports. their work is divided into three stages: (1) first, they fine-tuned the pre-trained Chex-Net to predict the specific tags from the images. (2) then calculated the weighted semantic features from the predicted tag's pre-trained embeddings. (3) finally conditioned a pre-trained GPT2 model on the visual and semantic features to generate the medical reports. They analyzed the generated reports using word-lapping metrics and adding new meaningful semantic-based similarity metrics. Similarly, Changchang Yin et.al [7] came up with a novel idea based on Hierarchical Recurrent Neural Network (HRNN) and introduced a novel frame to generate accurate and diverse medical reports. Their model can detect medical abnormalities and can generate long captions simultaneously. Moreover, they suggested replacing the global feature pooling in multi-label classification CNN with multi-label pooling to improve the accuracy and robustness of CNN.

Compared to earlier work, in this paper, we develop an algorithm that takes an X-ray image at the input and generates an automatic report at the output. In a nutshell, the algorithm combines Natural language processing and computer vision to accomplish the task. The paper is organized into the following sections. The methodology and important components of the model are elaborated in section 2. The data preparation and dataset details are listed in subsection 2.3. The experiments and implementation details are given in section 3 and section 4 concludes the paper with the final remarks.

2 METHODOLOGY

The proposed framework is inspired by Marc Tanti et.al [8] encoder-decoder architecture. We used a CNN encoder to extract visual features and an LSTM network to extract text features. The output of both CNN and LSTM is concatenated and given as input to a feed-forward network. The dataset used for this study is the IU chest radiology

The 18th Scandinavian Conference on Health informatics, Tromsø, Norway, August 22-24, 2022. Organized by UiT The Arctic University of Norway. Conference Proceedings published by Linköping University Electronic Press at https://doi.org/10.3384/ecp187.

dataset with chest images with their corresponding text reports. In our method, first, we extracted features from images using pre-trained models like VGG16, INCEPTION, RESNET, MOBILE NET, and NASNET MOBILE and saved those features into a dictionary and wrote it to a pickle file. The dictionary keys are the names of the images, and values for those keys were the features from the corresponding name. Then, we pre-processed the text data so the algorithm could process it. Then we passed the pre-processed text to the LSTM, where sequential features are learnt from the whole text, and then finally, the image features and the text features are concatenated in feed forward network.

Pre-processing

We have two modalities, i.e., images and text, so we pre-processed both data separately.

Each image in the dataset is resized to 500x500 pixels to reduce the computations and maintain the dataset's consistency. Then we normalized the images and scaled the image pixels to a range of [0, 255]. For the text data, to avoid inconsistency in the dataset, we converted all the letters in the reports to their lower case. Then we removed all the punctuation and special characters because we do not need that in our reports, and they could confuse our model while training. We removed those words from the reports that were repeated only once or consisted of a single letter, e.g., "a". As our desired output report does not contain any number, we checked if there were any alphanumeric characters and removed them too. To direct the model that our sequence starts and ends from here, we added "startseq" and "endseq" in the start and end of the reports, respectively. Then we created a vocabulary of the unique words in the reports. Then we tokenized the words to pass them to the model.

Feature Extraction

Several pre-trained models are used for extracting features like VGG16, GoogleNet, ResNet, and NasNet. We tested these models to determine which pre-trained model is best for report generation.

Dataset

For training the model, we used Indiana university's publicly available Chest X-Ray dataset [9] (IU X-Ray). The dataset contains 7,470 pairs of images and text reports. Each report consists of impressions, findings, tags, comparisons, and indications.

3 EXPERIMENTS

We implemented the model on a system with Core i5-6500 Processor, 24 GB RAM and GeForce GTX 1070 Ti Graphics card. We tested different backbones models on 80 epochs, including VGG16, Inceptionv3, Rasnet50, MobileNet and NasNet Mobile. The comparative analysis based on performance is tabulated in Table 1.

Models	Training Accuracy	Validation Accuracy	BLEU score
VGG16	95	93	0.94
Inception V3	95.6	88	0.93
ResNet 50	99.5	91	0.91
MobileNet	97	91	0.23
NasNet	99	93.8	0.11

Table 1. BLEU score, training, and validation accuracy on different backbone models.

4 SUMMARY

In this study, we proposed an automatic report generation system that generates radiology reports for chest x-rays using deep learning. The system uses CNN as encoder and LSTM as decoder for report generation. The experiment results show that the suggested system is reliable and fast. This reporting system not only has the potential to reduce radiology errors but also makes the radiology practice more efficient.

Reference

[1] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 721–729.

[2] Adrian Brady, Riste´ard ´O Laoide, Peter McCarthy, and Ronan McDermott, "Discrepancy and error in radiology: concepts, causes and consequences," The Ulster medical journal, vol. 81, no. 1, pp. 3, 2012.

[3] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[4] Yann LeCun, Yoshua Bengio, et al., "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, vol. 3361, no. 10, pp. 1995, 1995.

[5] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang, "Multimodal recurrent model with attention for automated radiology report generation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 457–466.

[6] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy, "Automated radiology report generation using conditioned transformers," Informatics in Medicine Unlocked, vol. 24, pp. 100557, 2021.

[7] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in 2019 IEEE international conference on data mining (ICDM). IEEE, 2019, pp. 728–737.

[8] Marc Tanti, Albert Gatt, and Kenneth P Camilleri, "Where to put the image in an image caption generator," Natural Language Engineering, vol. 24, no. 3, pp. 467–489, 2018.

[9] Dine Demner Fushman et al. "Preparing a collection of radiology examinations for distribution and retrieval," Journal of the American Medical Informatics Association, vol. 23, no. 2, pp. 304–310, 2016.