

Article

Transformer-Based Model for Radiology Text Reports Generation from Frontal and Lateral Chest X-ray Images

Abdelfettah Elaanba *, Mohammed Ridouani and Larbi Hassouni

RITM Laboratory, CED Engineering Sciences, Hassan II University, Casablanca 20430, Morocco;
mohammed.ridouani@gmail.com (M.R.); lhassouni@hotmail.com (L.H.)

* Correspondence author: abdefettah.elaanba@gmail.com

Received date: 5 March 2024; Accepted date: 21 March 2024; Published online: 10 July 2024

Abstract: A medical radiology report for a patient's chest X-ray image is a textual description of the abnormalities and normalities of the patient's lungs, heart, and chest. The report is a translated text according to the radiologist's diagnosis of the image. It helps to communicate the diagnosis to no radiologist or medical experts and track the patient's health condition in the future. In general, a medical report intended for use by busy medical practitioners in a stressful work environment requires a comprehensive report written with simple and grammatically correct language. The manual writing of reports is a tedious task that consumes time and is vulnerable to possible human errors. In this work, we propose a pipeline for building based-transformer models to process the task of automatic report generation. We included the experiments that justified the choice of the models and configurations of the feature extractor encoder, transformer attention layers, and training parameters. We show that (1) First, the pretrained ViT feature extractor performs better than the CNN-based encoder (DensNet121); and (2) Second, the dual-view input (Frontal and lateral images) give better result than one input (Frontal or lateral).

Keywords: transformers; radiology reports; CNN encoder; chest X-ray; vision transformer

1. Introduction

Radiology reports are the primary communication channel between radiologists or radiologists and other healthcare professionals. The report transfers results and helps provide solutions for health problems. The report is generally written manually by a radiologist. On average an expert radiologist spends more than 10 minutes [1] to write a radiology report. This issue consumes the expensive time of experts. Moreover, expert radiologist is considered a scarce resource. The stress caused by the workload and the task complexity can open a space for human errors. Besides those reasons an automatic radiology report generation system will save time and help in areas with fewer experts to automate the task, identify all possible normality or abnormality (normality or abnormality present in training dataset), and avoid the radiologist bias to look and search for a specific problem. In this work, we propose a pipeline to train transformer models as a step toward the development of such a system. The development of a general and efficient system for automatic radiology reports generation is retarded by several hump-stones. First, the dataset imbalance. The datasets contain reports and radiology images full of normality versus limited numbers of diseases (abnormalities) in boat images and text reports. Likewise, the available datasets suffer from the scarcity of diseases diversity, which prevent the generalization of report generation for all kind of health issues. Datasets also contains poor chest-X-ray image qualities, in addition X-ray images are very similar [2]. This similarity adds complexity to the models' training. The second hump-stone is the bias introduced by the sample studied and the methodology of radiologists and researchers in writing and processing a radiology report. The sample can concern a limited number of patients from one country, one city, or just one hospital sharing some comment living conditions. So, the results can't



generalize to the wider population. In addition, there is no standard methodology for the task of writing a radiology report. Besides this, a model trained in a specific study cannot directly applied to another context. The third problem is the nature of radiology reports with long and repetitive sentences, adds complexity to the problem compared to the normal captioning task requiring only the model to predict one sentence for an input image. Finally, the report's quality depends on the experience of the radiologist who wrote the report, which will influence the quality of reports generated by models trained on those reports. Although those reports are written by busy radiologists or sometimes by non-experienced doctors, the result is an unstructured report difficult to understand by the target practitioner or used to train an efficient model. Training a model to generate high-quality radiology reports, simple to understand by the busy practitioner, requires that the training dataset contain accurate and comprehensive reports with saint grammar and spelling [3]. For those reasons, we contribute with the trains of research to fix those issues, and we propose a flexible pipeline to train customized transformers for the text report prediction based on chest X-ray images. The pipeline will help to transcend the methodology and train models unto other private or public datasets for the given task of radiology report generation. As an application of the proposed method, we apply the pipeline using the IU-X-ray dataset. The IU dataset is a known benchmark used for X-ray reports generation related research evaluation and comparisons. In this paper, we employ the BLUE metric to evaluate our results. We present below the detailed descriptions for steps to use the proposed pipeline to customize and train a transformer model for similar tasks. This paper is structured as follows: Section 2 presents the related works to the automatic reports' generation task. Section 3, introduce the task of reports generation. In Section 4, we introduce the transformer models. Section 5 introduces the experiments. Section 6 presents the results and the analysis. Finally, in Section 7, we introduce the conclusion.

2. Related Works

Image captioning is the base-related task to radiology report generation. Gaung Li et al. [4] used a transformer for the image captioning task. The authors customized the attention and proposed an EnTangled Attention (ETA) to permit the model to capture semantic and visual information. The captioning consists of attributing one sentence to an image. It is a particular case for radiology reports with long paragraphs and multi-sentences. Each sentence describes a specific medical normality or abnormality for a precise zone in the image. Concerning chest X-ray images report generation Jianbo Yuan et al. [5] Propose an encoder-decoder model test on the IU dataset for multi-view chest X-ray images (Frontal and Lateral). The proposed encoder is a CNN model for feature extraction. The author tested with a pre-trained CNN on ImageNet and CheXpert datasets. The author reported that pre-training on domain-specific data performs better. For the decoder side, the author uses a hierarchical LSTM [6]. The author noticed that the proposed model performs like the state-of-the-art related baseline approaches. Using a different approach Omar et al. [7] proposed a conditioned transformer for radiology report generation. The authors fine-tuned a pre-trained CheXnet [8] for disease tag generation on the IU dataset. The tags were used later to fine-tune a pre-trained transformer GPT-2 for the task of report generation based on tags associated with the extracted decoder visual features from the Chest images. The authors report that the proposed solution performs more than transformers-based models in quantitative metrics. Moreover, the solution is faster to train. In the aforementioned work, the use of a large NLP pre-trained model like GPT-2, eliminates the need to specify a vocabulary for the model. In a different strategy instead of predicting the radiology report directly from the patient's chest image, Koji et al. [9] proposed a progressive framework starting with extracting global context text from the images in the first stage. Then, reforms the text with a transformer model (Text-to-Text) to generate the full radiology report. The authors report that the progressive approach surpasses standard and single Visual models in Blue-1,2,3 and ROUGE Metrics. The metrics measures were performed on two benchmarks: IU-X ray and MIMIC-CXR [10] datasets. Meanwhile, the model is under the baseline stat-of-the-art models in precision. Another two-stage architecture based on a reinforced transformer for medical report generation using the same IU dataset is proposed by Yuxuan et al. [11]. The authors proposed a visual decoder for visual feature extraction followed by a transformer decoder for the text report writing. The proposed model uses the self-critical reinforcement learning method. The authors report that the model improves the performance in the Blue-1 by 50% compared with the state-of-the-art image captioning models and augments the computation efficiency.

3. Radiology Reports Generation Task

3.1. Background

Given the use of attention mechanism by transformer models and their comparable or higher results on large image processing tasks compared to the CNN models, we choose to tackle the task of medical radiology images reports generation task with such models. The task of medical reports automatic generation is illustrated in the figure [Figure 1] below. Our objective is to propose a better architecture for a model that takes a patient X-ray image and outputs a medical report with an accurate diagnosis of this patient case. In this section, we start by presenting the report generation task, and the structure of the medical radiology report, then we will give an overview of sequence-to-sequence models. Their structure, training, and inference process.

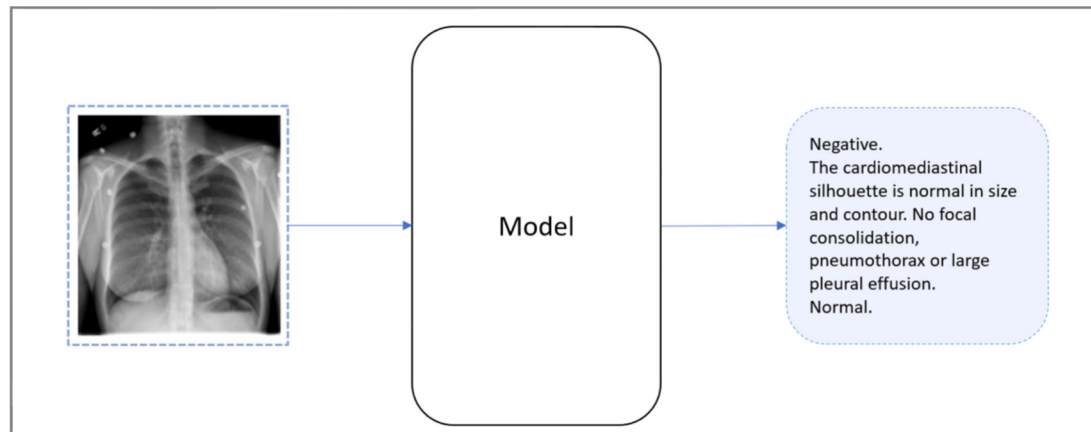


Figure 1. Medical Reports Automatic Generation Task.

3.1.1. Transformers vs. CNNs in Medical Images Applications

Kelie et al. in the paper ‘Transformers in medical image analysis’ [12] report a comparison between transformers and CNNs models [13–19] in medical image tasks. The authors mention that transformers perform like CNNs or better within most tasks. Second, the transformers perform better on large-scale datasets. In contrast, for the weaknesses, the authors mention the larger computation cost of transformer training. This issue can be fixed if we optimize the models’ architectures. Finally, the authors noticed that a hybrid model transformer-CNN has the advantages of boat attention and convolution combined. The attention in the transformer models used to capture the long-term dependencies, and convolution in the Convolutional Neural Networks captures the short-term dependencies.

3.1.2. Transformers Applications in Medical Image Analysis

The success of transformers with NLP tasks transcends recently to vision tasks. Medical image analysis is not an exception. Kelie et al. [12] analyses the specter of the model’s application in the medical field and names the tasks with the used transformer models architectures. The authors named a list of applications; classification, segmentation, registration, video-based application, detection, image denoising, image synthesis, and image super-resolution. The authors compare transformer models to other deep learning models (CNNs) and shows that transformers perform better on most of the tasks. In addition, the authors compare different transformer architectures for each task and show which fit each one.

3.2. Radiology Report Structure

A medical report is a text document written by a radiologist to diagnose a patient’s health condition based on the radiologist visual observations [20] while paying attention to the patient alternative diagnosis modality and his health history records. The report’s audience is other radiologists, doctors, patients, and their families. A chest X-ray report is a semantic description of one or several chest X-ray images using text. The report is composed of a secession of sentences. Each sentence highlights an abnormality or a normality. The length of a report can be long (more than 1000 words) or short, depending on the problems present in the X-ray image of a given patient. Knowing that the primary audience for a radiology report is the busy clinician with limited time who needs to read and understand the report rapidly without any confusion. Besides this, the report should be logically organized and

written with clean language, good grammar, and spelling [3]. The quality of the automatic report generated by a trained model will depend on the quality of radiologists' written report present on the training data set. So, the first stage to developing an efficient model is to have access to high-quality data. The report is subdivided into three parts. (1) impression which serves as a report title. (2) the finding contains the full report, and (3) the manual tags which consist of a keywords list used as a report summary. Depending on the report format a recommendation or conclusion paragraph can included and serve as a guide to show how the report is understood [21]. In general, report writing follows a known pattern but there is no standardized unique format to write the report [22]. The image Figure 2 below shows an example of a chest X-ray report from the IU-Dataset [23].

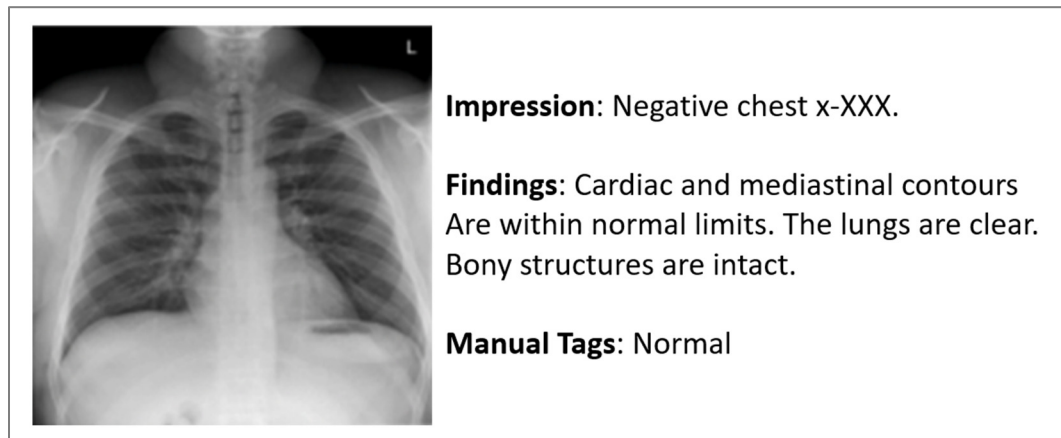


Figure 2. A sample image from the IU-Chest X-ray dataset [23].

3.3. Radiology Report Generation Methods

In typical work conditions, a radiology report is written by a radiologist within an average of 10 min. The radiologist looks at the patient's X-ray images to deduce abnormality or normality to report. In general, a radiologist doesn't study how to write a typical report [24]. So, the practice and experiment are the way to produce the report text. The consequence of this, is that reports are not standardized. From a deep learning view, it is difficult to train models when data is not homogenous. In practice, the dataset for the task of automatic report generation is annotated by a group of radiologists with a defined and unified methodology. The second method to resolve the problem is with the application of deep-learning Image text model's, in particular transformer models which is the object of this paper. The principle is to use supervised learning to train a model on X-ray images with reports as a label and use the Gradient to minimize the loss function. We will detail the pipeline of the training and inference in the next few sections.

4. Sequence to Sequence Models

In deep learning, a sequence-to-sequence model takes a bench of tokens input and outputs a sequence of tokens. The input and output can be homogeneous (Text to Text, Image to image, speech to speech.) or heterogeneous (text to image, image to speech...). If we consider an input $X(x_1, x_2, x_3, \dots, x_n)$ and $Y(y_1, y_2, \dots, y_n)$, the output sequence of the model can be represented as a function $F(X) = Y$. The function F represents the task performed by the S-T-S model. This task can be translation, image captioning, or text summarization to name a few. F is predicted by the minimization of the loss function during training using the stochastic gradient descent optimization [25]. Such models are used first for natural language processing such as translation. The task consists of taking a words sequence from a source language and seeking to get the corresponding output using a target language. One of the used models is RNNs [26].

4.1. RNNs Models

RNN is a recurrent neural network used for sequential data. If we take the natural language processing as an example, the RNN model takes the input sequence word by word to generate the output word by word. The model is composed of two components: An Encoder and a Decoder. The encoder is a succession of RNN cells with a token as an input associated with the previous cell output to generate a new hidden state. The process is repeated as we have a word in each sentence. The final cell outputs a context vector that englobes the meaning and the relation between words in the full input sentence. The context vector is passed to the decoder to begin the decoding process word by word until the end of the

task (Translation as an example). Each step in the decoder provides a word used as input for the next step. For this example of using RNN for NLP processing, the model type is multi-word input and multi-word output, known as many-to-many architectures. There are other types of RNNs used for different tasks; The many-to-one architecture is used for movie sentiment detection, and the one-to-many architecture is used for image captioning tasks. The diagram below shows the different possible architectures of the RNN model Figure 3.

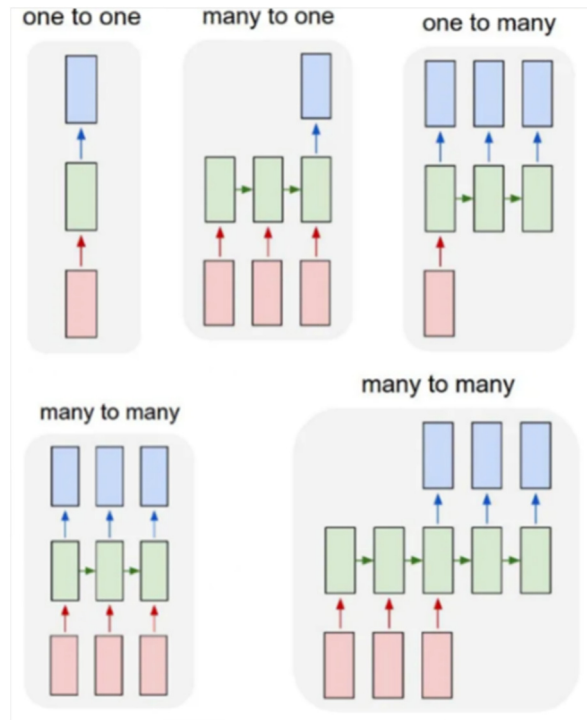


Figure 3. RNNs possible input-output architectures.

4.2. Transformers

Because of the recurrent property of RNNs, the long path can cause the loss of relevant information along the way (gradient vanishing). Moreover, the calculation cannot be paralleled. To fix this problem, an attention mechanism is introduced to give the decoder the possibility to look at all encoder hidden states. This fixes the issue of information loss. For the parallel calculation, the recurrent connection is dropped giving the encoder the ability of independent processing. Besides the absence of the recurrence in transformers, the sequence order is not tracked in the transformer models. This problem is mitigated by adding a position encoding layer for the transformer’s encoder input. Inspired by RNNs and attention mechanism, transformer models were first proposed by Google in the famous paper ‘Attention is all you need’ [27]. The model in the paper is designed to fix the common issues in NLP models. The authors also mentioned that the model can be good at other tasks like image processing. Which was experimented with and proved a few years after the big success of NLP. The Transformer models become the gold standard in computer vision by 2020. The transformers are maintaining the Encoder-decoder architecture Figure 4 enhanced by the attention mechanism. The name of the transformer came from the action made by the attention layer, transforming a given sequence to an output sequence. We will briefly describe the transformer architecture and blocks in the paragraph below.

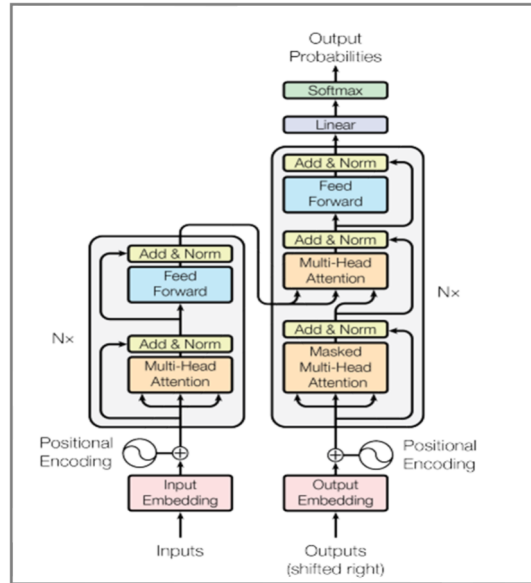


Figure 4. Transformer model architecture [27].

4.2.1. Attention Mechanism

Inspired by the human ability to process real time data stream by using attention to focus on relevant information and ignore the unimportant parts. The attention mechanism was first introduced by [28] as “Bahdanau Attention”, which is applied at the beginning for the language-translation task. The attention is used to find the dependencies between the elements of an input sequence (Sentences in NLP). Two attention types persist: First, the self-attention where this dependency is computed between the encoder input sequence elements (Intra-sequence dependency). Second, the attention where the dependence computed between the encoder output sequence and the decoder input sequence (Inter-sequence dependency).

The attention is calculated as a vectorial projection of a transformed input vector X (weighted by the encoder weights) into three vectors Q , V , and K using the equation below. The Q is the query vector $Q = W_q * X$. K is the Key Vector $K = W_k * X$, and V is the Value vector $V = W_v * X$.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The difference when computing the self-attention and the attention is; for the first, the V , K , and Q derived from the same encoder input X . Otherwise, for the second, the V and K derived from the encoder source input X , while the Q derived from, the target, the decoder input sequence.

4.2.2. Transformer Encoder and Transformer Decoder

The encoder’s role is to find the dependence in the input sequence known as the context vector. The context vector is transferred to the decoder and used to compute the inter-attention with the target output sequence. An encoder or a decoder is a pile of N transformer blocks, except the encoder only uses self-attention. The decoder uses both self-attention and attention. A transformer bloc contains an embedding layer followed by a position encoding layer (only the first block in the encoder-decoder). Next, follow a multi-head self-attention layer and a normalization layer. Another attention layer is added only for the decoder blocks. Then, a feed forward and a normalization layer at the top of the encoder-decoder blocks. The output of the transformers model is a SoftMax activation applied on the last block of the decoder.

4.2.3. Tokenization

A deep learning model cannot manipulate alphabetic characters directly. The loss function minimization during training needs numbers to compute the gradient that gives insight into where the direction moving model weights to reach a minimum. Besides this, the words are embedded vectors before the transformer input layer. This operation is known as tokenization. In NLP, there are three types of tokenization: Word-level tokenization, character tokenization, and sub-word tokenization. Each tokenization type has advantages and inconveniences. The word level type is fast but not flexible for typos. Otherwise, the character tokenization needs much computation to learn known words from

character representation. Last, but not least, sub-word tokenization is the more convenient method. Sub-word tokenization, which is Middle Ground tokenization, requires around 30 K tokens to encode any sequence, even non-alphabetical ones.

4.2.4. Position Encodings

Unlike other sequential models such as RNN, for which the sequences processing order is explicitly defined by the recurrence nature of the model (First word processed is the first in the sentence and so on), the transformer models process in parallel and prevent tracking time or order of sequences. To add order which is important in NLP (order of words in a sentence) or patches in an image (In CNN models the position is tracked by the convolutions operations), a position vector is added to the sequence vector after the input embedding layer for both encoder and decoder for a transformer model. In the original transformer paper ‘Attention Is All You Need’, a continuous sinusoidal function is used to encode the position. The continuousness of the encoding is better than a simple one-hot vector position encoding in the sense of the model’s ability to input sequences longer than ones encountered during training.

4.2.5. Transformer Training and Inference

During training, the transformer has access to labeled data for the desirable task. For example, if we consider the translation task, we feed the encoder by the source sequence, and the decoder by the target sequence. Then, we measure the loss according to the transformer’s ability to predict the next word into a translated sentence. Next, we update the encoder-decoder weights using gradient descent back-forward propagation. The training process changes slightly according to the learnable task. Another example is text generation. If we take the famous model GPT-2, an encoder with 48 transformer blocks, the model is trained to guess the next word in a sentence. The training was unsupervised on a large high-quality internet text. For a random sentence, the next word is masked and used as a label, and the model is trained to predict this next word (Masked word in the original sentence). Otherwise, during inference in the case of an encoder-decoder transformer, we fed only the encoder by the input sentence (source), and the decoder progressively auto-alimented by the predicted token from the previous step until the complete target sentence is generated.

4.2.6. Vision Transformers

After the transformer’s success in natural language processing tasks, the transformer’s models use transcends to broader tasks like computer vision. Besides the transformer being a sequence-to-sequence models, an input image is split into patches, flattened to one sequence vector, and next, a position attributed to each patch (position encoding). The processing in the other layers is like NLP transformers. The reason tokens in a Vision transformer ViT [29] are patches, not pixels, is to reduce computation cost into the attention layer which is the square of the token number. For example, an image with the shape of $(N, N, 1)$ requires a multiple of $N*N$ operations for one attention head (N is a natural number). Patch coding for image processing task is equivalent to using sub-word tokenization instead of word or character tokenization in NLP processing with transformer model to reduce computational cost.

5. Experiments

5.1. Dataset Presentation

IU-dataset is considered one of the few benchmarks publicly available to evaluate the automatic reports generation task. In this work we used the IU dataset to test our transformer models and compare our results. The Indian Dataset is available for research and downloading from the National Library of Medicine (<https://openi.nlm.nih.gov/> accessed on: 20 August 2023). The dataset contains 7,470 Chest X-ray images (Frontal and Lateral images) collected from Indiana Network for Patient Care. With 3,955 corresponding radiology reports. Each report in the dataset is associated with two chest X-ray images, one frontal and a lateral view. To preserve data privacy, de-identification is applied to all reports and DICOM images. This operation removes any information that can identify a patient’s identity. IU-dataset is one of the few public data sets with chest X-ray reports. However, it’s considered one of the best available benchmarks to compare results in tasks like chest-X ray medical reports generation task.

5.2. Data Analysis and Preprocessing

For the dataset preparation, first we randomly split the data into train, validation, and test (70%, 10%, 20%). The split percentage mimics the dataset preparation in the related works to have a similar benchmark to compare our results because there is no standard evaluation procedure for the report

generation task. For the report text preprocessing, we convert all characters to lowercase. Then, we remove no alphanumeric characters. Next, we tokenize the text.

5.3. Metrics for Evaluating Text Generated Quality

Evaluating how good a model in the chest X-ray medical reports generation is to find a metric to measure predicted and original report resemblance. The idea is to find a metric that compares on a semantic level the two reports. This kind of metric can be calculated only by humans looking manually and comparing the reports. In practice, there are several metrics that look to exact or synonym words into source and target text. An example is the BLEU [30] metric which calculates the number of exact words and gives a score between 0 and 1. This is the Blue-1 metric. There are other variants of BLUE that track the number of identical sequences with N successive words in the compared reports named BLUE-N. N can be an integer number > 0 . In addition, two metrics designed first for language translation are ROUGE-L [31] and METEOR [32]. Practically ROUGE-L metric is used for the estimation of text summary quality. Last, but not least the CIDEr [33] metric is used to evaluate the image captioning task.

5.4. Methodology

For this work, we propose to use a transformer model for the task of medical images reports generation. The proposed architecture is formed with a feature extractor model followed by an encoded-decoder transformer. For the Input, we train our model using a single (Frontal image) or dual-view input (Frontal and lateral chest X-ray images). The model architecture is reported in the diagram below Figure 5.

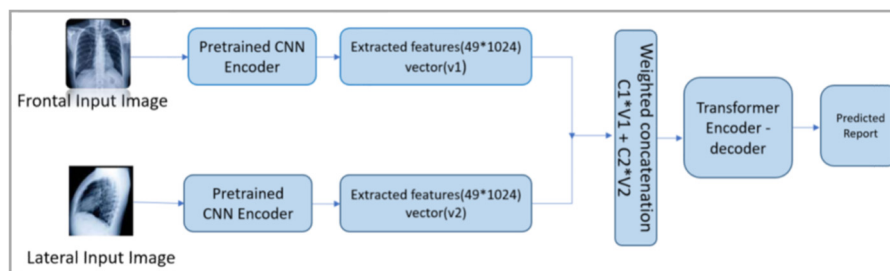


Figure 5. Multi-inputs transformer architecture.

In the case of double inputs, we feed the model with two images for the same patient (frontal and lateral views). Each image is processed by a separate pre-trained CNN encoder. Then, the two output vectors are concatenated to form a single vector. The resulting vector is fed through the transformer encoder input to generate the final report.

Otherwise, for the standard architectures with one image as input Figure 6, we use the frontal image to train our transformer model. The training and inference processes, in this case, are like the previous dual-inputs framework with only one CNN Encoder instance.

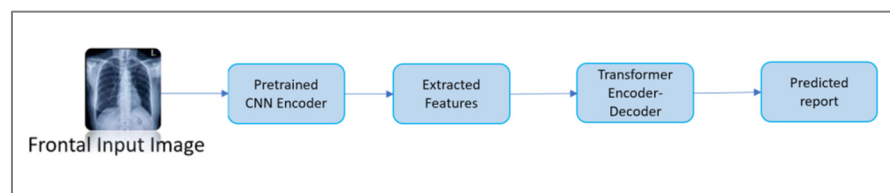


Figure 6. Single input transformer architecture.

For the training and the test procedures, we train our model variants on the training data set by minimizing the loss function. Then, we compute the BLUE-1 metric on the test dataset. The objective is to compare the efficiency of different feature extractor encoders and different transformer model variants. The variant differs by the number of attention layers and the number of heads per attention layers. For the training hyperparameters, we test different configurations for the learning rate, batch size, input images, and optimizers. The best parameters are reported in the table below Table 1.

Table 1. Training Hyperparameters.

Training Hyper Parameters	
Initial Learning rate	1×10^{-4}
Image Input	(224,224,3)
Augmentations	RandomContrast, RandomRotation
Batch size	16
Loss function	SparseCategoricalCrossentropy

6. Results and Analyses

The table [Table 2] below shows the results of different tested features extractor encoders with our proposed transformer model. The test aims to compare the pre-trained cnn-encoders with pre-trained ViT Models. After the training using the same transformer model configurations, we noticed that the ViT encoder trained on imageNet performs better than all tested CNN-based encoders. Furthermore, the CNN-encoders pre-trained on the imageNet dataset outperform the one trained on the chest X-ray 8 dataset. Besides the chest X-ray 8 contains chest images similar to those of the processed problem in this paper, we aim to get a higher result than the training on imageNet containing images for general context. However, the experimental result shows that the Densnet121 encoder pre-trained on ImageNet (BLUE = 0.15) outperforms the pre-trained Resnet121 on the Chest X-ray 8 database (BLUE = 0.13). We explain these results by the fact, that even if the Chest X-ray-8 data set contains X-rays images, the diseases processed with this dataset are different from those processed in the IU dataset. So, the pertained model on ImageNet is more general and can be used for the specific task of X-rays report generation. We know that transfer learning [34] is efficient from a general to specific tasks not between two specific tasks. The transfer of learning is effective when the two specific tasks are similar, which is not the case in this given scenario.

Table 2. Features extraction Models.

Features Extraction Models	BLUE-1
ViT-Encoder (Image Net)	0.26
InceptionV3 (ImageNet)	0.19
DensNet121 (ImageNet)	0.15
DensNet121 (Chest X-ray 8)	0.13

After the selection of the feature extractor encoder model, we fix the chosen model in our architecture, and we start testing different transformer configurations for the encoder-decoder model's. The table below shows the results of our tests Table 3. For each test, we change the number of blocks and the number of heads per block. The 3 blocks with 16-head transformer model perform better than all tested versions.

Table 3. Transformer models configurations tests.

Transformer Model	Number of Attention Layers	Number of Heads per Attention Layer	BLUE-1
Transformer1-2	1 layer	2 heads	0.13
Transformer1-4	1 layer	4 heads	0.15
Transformer1-8	1 layer	8 heads	0.22
Transformer1-16	1 layer	16 heads	0.23
Transformer2-16	2 layers	16 heads	0.23
Transformer3-16	3 layers	16 heads	0.26

The result shows that the precision augments broadly for the one attention bloc transformer model when we augment the number of heads (2, 4 and 8 blocs). The improvement became small with big number of heads and more blocs (transformer1-16, transformer2-16, and transformer3-16). The explanation for this behavior is that with the larger models, the training tends to overfit because of the small size of the training data set (IU-dataset) used in this work.

After the selection of the best model from the tests above. We proceed to test the model using our proposed architecture. First, with two images as input (frontal and lateral images). The input is the sum of the CNN encoder extracted features for the frontal (F1) and the lateral image (F2). To simplify the test, we suppose that the frontal and the lateral images contribute equally to the result [eq]. Then we compare the results obtained with a single image input (Frontal or Lateral).

$$\text{Input} = C1*F1 + C2*F2 \quad (C1=C2=1) \quad (2)$$

The table [Table 4] below shows the result for different possible inputs (Frontal only, Lateral only, weighted sum of lateral and frontal images).

Table 4. Weighted sum input tests.

Model Input	BLUE-1
Frontal images	0.26
Lateral Images	0.23
Weighted feature sum F1 + F2	0.27

The table shows that dual input performs better than a single input image (frontal or lateral X-ray images). This portends that the combination of two inputs gives the model the possibility to find additional information that is present in one image and not in the other. Or to reinforce the redundant information found in both inputs. Below are some sample reports generated by our model Figure 7. We noticed that the majority of reports were filled with normalities rather than anomalies. Additionally, the blue score for normalities is higher, this is due to the imbalance of the training data set.






	Blue: 0.29	<p>Predicted Report: heart size and mediastinal contours are within normal limits the lungs are clear no focal airspace consolidation no pleural effusion or pneumothorax is seen</p> <p>Reel Report: Heart size within normal limits. No focal airspace disease. No pneumothorax or effusions.</p>
	Blue: 0.57	<p>Predicted Report: the heart is normal in size the mediastinum is unremarkable the lungs are clear.</p> <p>Reel Report: The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.</p>
	Blue: 0.12	<p>Predicted Report: the heart is normal in size the mediastinum is unremarkable the lungs are clear there is no pleural effusion or pneumothorax the right upper lobe there is no acute bony structures are unremarkable.</p> <p>Reel Report: Both lungs are clear and expanded. Heart and mediastinum normal.</p>
	Blue: 0.60	<p>Predicted Report: the heart and lungs have xxxx xxxx in the interval both lungs are clear and expanded heart and mediastinum normal</p> <p>Reel Report: The heart and lungs have XXXX XXXX in the interval. Both lungs are clear and expanded. Heart and mediastinum normal.</p>
	Blue: 0.10	<p>Predicted Report: the heart is normal in size the mediastinum is unremarkable the lungs are clear there is no pleural effusion or pneumothorax the right upper lobe there is no acute bony structures are unremarkable</p> <p>Reel Report: The cardiac contours are normal. The lungs are clear. Thoracic spondylosis.</p>

Figure 7. Reports Samples Generated By our Model.

7. Conclusions

The proposed pipeline is a step toward the development of an efficient automatic system for medical report generation task based on chest X-ray inputs. Such a system will considerably reduce the load work charge of radiologists and help them inspect the outputs of the system to reduce possible diagnosis errors. We proposed the use of attention-based models such as transformers, which are state-of-the-art captioning models. The development of the report generation task needs more annotated data and a global standardization for how to write reports. We consider for our future works to experiment with the coefficients of the dual view input to find the exact contribution of the frontal and the lateral inputs in the result. Finally, we assume that developing an effective medical reporting system requires a two-sided solution; the first side is an effective model, and the second is the quality of data annotation which depends on the expertise of the radiologist and the time spent on the annotation task.

Author Contributions

A.E., M.R. and L.H. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Data Availability Statement

The Dataset used to train our models is available for research and downloading from the National Library of Medicine.

References

1. I.A., Cowan, S.L.S., MacDonald, R.A., Floyd. Measuring and managing radiologist workload: measuring radiologist reporting times using data from a Radiology Information System. *J. Med. Imaging Radiat. Oncol.* 2013 Oct;57(5):558-66. doi: 10.1111/1754-9485.12092. Epub 2013 Jul 12. PMID: 24119269.
2. Wang, Z., Han, H., Wang, L., Li, X., Zhou, L. Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach. *IEEE Trans Med Imaging* 2022 Oct;41(10):2803-2813. doi: 10.1109/TMI.2022.3171661. Epub 2022 Sep 30. PMID: 35507620.
3. A. Lukaszewicz, J. Uricchio, G. Gerasymchuk. The art of the radiology report: practical and stylistic guidelines for perfecting the conveyance of imaging findings. *Can. Assoc. Radiol. J.* 2016; 67: 318-21.3.
4. G. Li, L. Zhu, P. Liu and Y. Yang, "Entangled Transformer for Image Captioning," In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019*, pp. 8927-8936, doi: 10.1109/ICCV.2019.00902.
5. J. Yuan, H. Liao, R. Luo and J. Luo (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22* (pp. 721-729). Springer International Publishing.
6. A. Graves and J. Schmidhuber (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, 18(5-6), 602-610.
7. O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, A. Fahmy (2021). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24, 100557.
8. P. Rajpurkar, J. Irvin, K. Zhu, et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv arXiv:1711.05225*.
9. F. Nooralahzadeh, N. P. Gonzalez, T. Frauenfelder, K. Fujimoto and M. Krauthammer (2021). Progressive transformer-based generation of radiology reports. *arXiv arXiv:2102.09777*.
10. A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C. Deng, R.G. Mark. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1), 317.
11. Y. Xiong, B. Du, P. Yan (2019). Reinforced Transformer for Medical Image Captioning. In *Machine Learning in Medical Imaging. MLMI 2019*, Suk, H.I., Liu, M., Yan, P., Lian, C. (Eds). *Lecture Notes in Computer Science*, vol 11861. Springer, Cham. https://doi.org/10.1007/978-3-030-32692-0_77
12. K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang. (2023). Transformers in medical image analysis. *Intelligent Medicine*, 3(1), 59-78.
13. J. Rabbah, M. Ridouani, L. Hassouni. A New Classification Model Based on Transfer Learning of DCNN and Stacknet for Fast Classification of Pneumonia through X-ray Images. *International Journal of Reliable and Quality E-Healthcare*, 2023, 12(1).
14. K. Elaissaoui, M. Ridouani. Application of Deep Learning in Healthcare: A Survey on Brain Tumor Detection International Conference on Connected Object and Artificial Intelligence (COCIA'2023). *ITM Web Conf.* Volume 52, 2023.

15. A. Elaanba, M. Ridouani, L. Hassouni. Automatic Diagnosis Chest-X-ray-Based-Framework for Semantic Segmentation and Placement Errors Detection of Catheters and Tubes. *International Journal of Computer Information Systems and Industrial Management Applications* 2023, 15(2023), pp. 257–265.
16. C. Techa, M. Ridouani, L. Hassouni, H. Anoun (2023). Automated Alzheimer's Disease Classification from Brain MRI Scans Using ConvNeXt and Ensemble of Machine Learning Classifiers. In: Abraham, A., Hanne, T., Gandhi, N., Manghirmalani Mishra, P., Bajaj, A., Siarry, P. (Eds). In *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)*. SoCPaR 2022. *Lecture Notes in Networks and Systems*, vol 648. Springer, Cham. https://doi.org/10.1007/978-3-031-27524-1_36.
17. A. Elaanba, M. Ridouani, L. Hassouni (2023). Automatic Diagnosis Framework for Catheters and Tubes Semantic Segmentation and Placement Errors Detection. In *Innovations in Bio-Inspired Computing and Applications*. IBICA 2022, Abraham, A., Bajaj, A., Gandhi, N., Madureira, A.M., Kahraman, C. (Eds). *Lecture Notes in Networks and Systems*, vol 649. Springer, Cham. https://doi.org/10.1007/978-3-031-27499-2_17.
18. C. Techa, M. Ridouani, L. Hassouni, H. Anoun (2023). Alzheimer's Disease Multi-class Classification Model Based on CNN and StackNet Using Brain MRI Data. In: Hassanien, A.E., Snášel, V., Tang, M., Sung, T.W., Chang, K.C. (Eds). In *Proceedings of the 8th International Conference on Advanced Intelligent Systems and Informatics 2022*. AISI 2022. *Lecture Notes on Data Engineering and Communications Technologies*, vol 152. Springer, Cham. https://doi.org/10.1007/978-3-031-20601-6_23.
19. A. Elaanba, M. Ridouani, L. Hassouni. A Stacked Generalization Chest-X-ray-Based Framework for Mispositioned Medical Tubes and Catheters Detection. *Biomedical Signal Processing and Control* 2023, 79, 104111.
20. M. Pahadia, S. Khurana, H. Geha, and S.T.I. Deahl. "Radiology report writing skills: A linguistic and technical guide for early-career oral and maxillofacial radiologists," *Imaging Sci Dent*, vol. 50, no. 3, p. 269, 2020, doi: 10.5624/isd.2020.50.3.269.
21. Michael P. Hartung, Ian C. Bickle, Frank Gaillard and Jeffrey P. Kanne. 2020 by the Radiological Society of North America, Inc. <https://doi.org/10.1148/rg.2020200020>.
22. A.P. Brady Radiology reporting—from Hemingway to HAL? *Insights Imaging* 9, 237–246 (2018). <https://doi.org/10.1007/s13244-018-0596-3>.
23. D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani. Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA* 23(2), 304–310 (2016).
24. A. Wallis and P. McCoubrie. The radiology report—are we getting the message across? *Clin Radiol*. 2011 Nov;66(11):1015-22. doi: 10.1016/j.crad.2011.05.013. Epub 2011 Jul 23. PMID: 21788016.
25. S. Ruder (2016). An overview of gradient descent optimization algorithms. *arXiv arXiv:1609.04747*.
26. D.E. Rumelhart, G.E. Hinton, R.J. Williams (Sept. 1985). Learning internal representations by error propagation. Tech. rep. ICS 8504. San Diego, California: Institute for Cognitive Science, University of California.
27. A. Vaswani, N. Shazeer, N. Parmar (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30.
28. D. Bahdanau, K. Cho, Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate 2014. *arXiv arxiv:1409.0473*. doi:10.48550/ARXIV.1409.0473
29. L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.H. Jiang, F.E.H. Tay, J. Feng, S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2021.
30. K. Papineni, S. Roukos, T. Ward, W.J. Zhu. "Bleu: a method for automatic evaluation of machine translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 2002.
31. C.Y. Lin. "Rouge: A package for automatic evaluation of summaries." In *Text Summarization Branches Out* 2004.
32. S. Banerjee and L. Alon. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 2005.
33. R. Vedantam, C. Lawrence Zitnick and P. Devi "Cider: Consensus-based image description evaluation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2015.
34. K. Weiss, T.M. Khoshgoftaar, D.D. Wang. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 1–40.