

# Radiology Report Generation using Full Transformer Architecture

**Mohammad Sabik Irbaz**

ID: 160041004

sabikirbaz@iut-dhaka.edu

**Abir Azad**

ID: 160041024

abirazad@iut-dhaka.edu

## Abstract

Radiology report generation systems can offer the potential to accelerate clinical processes by saving radiologists from the repetitive labor of drafting radiology reports and preventing potential minor and major medical errors. It is considered one of the hardest tasks in medical domain. But due to recent revolution brought by Transformer architecture, researchers are looking into image-to-text transformation methods to solve this problem. We propose our novel approach which uses a full transformer architecture which reached the previous state-of-the-art on MIMIC-CXR dataset even when we trained with a smaller subset of the whole dataset.

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Literature Review</b>	<b>3</b>
2.1	Word Embeddings . . . . .	3
2.2	Seq2Seq . . . . .	4
2.3	Attention . . . . .	4
2.4	Transformer . . . . .	4
2.5	Vision Transformer (ViT) . . . . .	5
<b>3</b>	<b>Related Works</b>	<b>5</b>
<b>4</b>	<b>Proposed Methodology</b>	<b>6</b>
4.1	Extracting and encoding image features . . . . .	6
4.2	Decoding the encoded images with transformer decoder . . . . .	7
4.3	Cross-Attention . . . . .	7
<b>5</b>	<b>Experimental Analysis</b>	<b>7</b>
5.1	Dataset . . . . .	7
5.2	Experimental Setup . . . . .	7
5.3	Result Analysis . . . . .	8
<b>6</b>	<b>Conclusion and Future Plans</b>	<b>8</b>

## List of Figures

1	Examples of Radiology Report Generation task. . . . .	3
2	Encoder Decoder Model . . . . .	4
3	Attention Mechanism . . . . .	5
4	Vision Transformer . . . . .	6

5	Transformer achitecture . . . . .	7
6	Full Transformer Architecture (Ours) . . . . .	8

**List of Tables**

1	Comparative result analysis . . . . .	8
---	---------------------------------------	---

## 1 Introduction

Researchers from all-over the world have been trying to automate the process of radiology report generation from X-Ray images since the task have been very cumbersome for the radiologists and there are always potential chances of human errors. Before 2019, the main barrier was that there was not a big enough trust-worthy dataset. In 2019, MIMIC-CXR (Johnson et al., 2019) was released which was annotated and curated mainly for this task. Even after that, not much work has been done with this since the task is, by nature, very complex. All the previous works used a CNN+RNN type architecture which actually had some bottleneck issues.



Images	Reference
	PA and lateral views of the chest are obtained. There is <u>mild atelectasis at the left lung base</u> . The previously seen endotracheal tube and nasogastric tube are no longer present on this study. There is no evidence of pneumonia, pleural effusion or pulmonary edema. The cardiomediastinal silhouette is unremarkable.
Images	Reference
	There is <u>moderate pulmonary edema</u> , but no pleural effusion or pneumothorax. Heart size is top-normal, stable. Mediastinal contours are within normal limits. Osseous structures are intact.

Figure 1: Examples of Radiology Report Generation task.

We proposed a new architecture which utilizes the transformer architecture. This type of architecture is one of a kind for this task and in this domain. The invention of vision transformer made our task a lot easier. The main challenges we faced is to fine-tune two pre-trained models which are trained on different condition and obviously different year.

To achieve SOTA performance, we still need to do more work. The results we got till now are promising. Further experimentation needed before reaching any conclusion.

## 2 Literature Review

### 2.1 Word Embeddings

In short Word Embeddings are numerical representations of text. As human we can percieve languages and realize its meaning. We know the difference between "king" or "queen" and "man" or "woman". But for computers its not so much evident. The idea is that we create a high dimensional vector space where we can represent each word using a unique vector representation, where the similar words will have similar representation or more specifically they will be closer in the vector space. Implementation of word embedding brought about a revolution in natural language processing as it could now effectively detect the similarity or difference between words within sentences. As they are represented in vector form, linear operations could easily modify

the word representations. One perfect example would be like, "King" - "Man" + "Woman" = "Queen".

## 2.2 Seq2Seq

Seq2Seq or Sequence to Sequence modeling was first proposed by Google (Sutskever et al., 2014), a general end-to-end approach for sequence learning. This was primarily introduced for neural machine translation, translating from one language to another language. Later it started become popular in text summarization, Sentiment conversion and all types of sequential models. Sentences are basically a sequence of words that can be converted into a sequence of word vectors or embedding. From the perspective of architecture, Seq2Seq models have two parts- the Encoder and the Decoder. (Cho et al., 2014) The input sequence goes through the Encoders sequentially. The encoders are basically a series of RNNs or LSTMs. Each input vector goes through an encoder RNN and forwards a hidden representation to the next layer along with the next input vector. Thus each output of the RNN contain all the information of the previous input sequences. Thus final a context vector is created which is called Encoder representation. Now the model need to decode the context vector to the preferred output sequence through the Decoder which is also a series of RNNs or LSTMs. The decoder RNNs decode the context vector sequentially by producing one output vector at a time. Figure 2 shows an example of Encoder-Decoder based Seq2Seq model.

## 2.3 Attention

In the Encoder Decoder model the sequence data is computed one by one. Due to its long correlated inter-dependencies in the context vector, for larger sentences it is more likely that the initial information might be lost. That's why attention mechanism was introduced. In the figure the encoder decoder model with attention is shown. Here the words 'How', 'are', and 'you' are converted into word embedding in creating the encoder vector. While decoding, apart from the context vector, the decoder is also fed with a direct connection from the input sequence representing the relevance of the corresponding words, dictating where the model should focus more seriously to predict the generated sentence precisely. Figure 2 shows sequence to sequence model without attention and Figure 4 shows sequence to sequence model with attention.

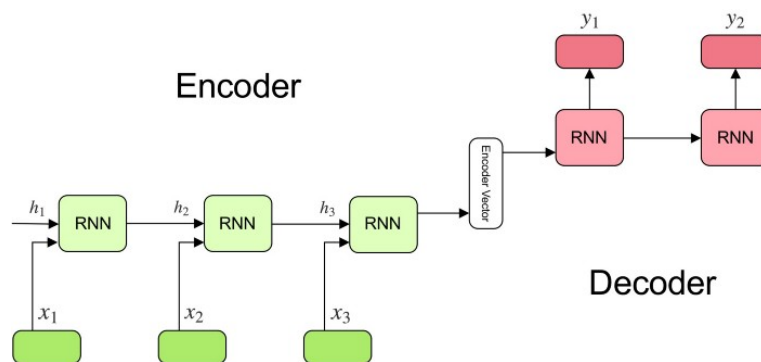


Figure 2: Encoder Decoder Model

## 2.4 Transformer

In the Transformer model, (Vaswani et al., 2017) proposed a new network architecture for sequence to sequence modeling solely based on the attention mechanism. Sequential models like RNN and LSTM process the sentences word by word, thus creating a huge problem in the process of parallelization. They introduced multi-head attention in their architecture. It's

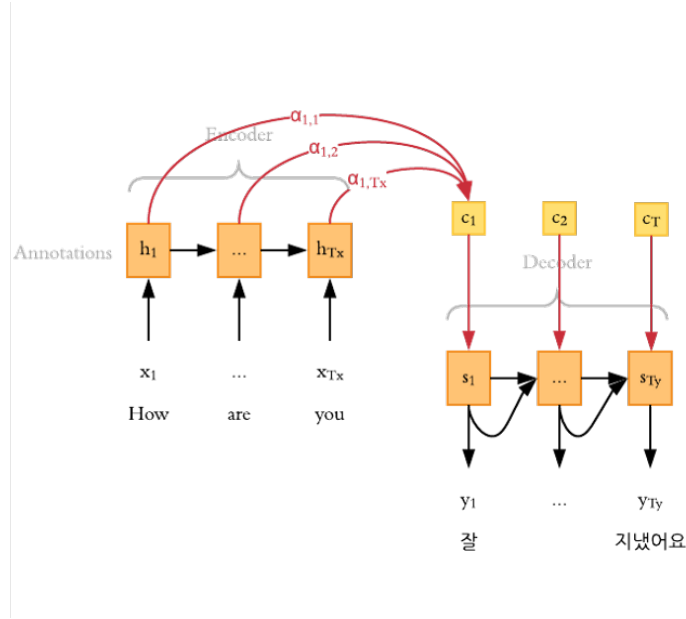


Figure 3: Attention Mechanism

basically the implementation of self attention, where the input sequence learns the similarities among itself denoting which words more related to each word. Thus it can have a deeper understanding about the language and its construction. The multi-headed attention is learned using three vector values, Q (query), K (key) and V (value). These parameters essentially contain the relationship among the word vectors. One of the reasons sequential models like RNN were so successful for language processing, was that it actually goes through the sentence one by one, thus learning the positional relationship among the words. For solving this issue in Transformer, Positional Encoding was introduced. It is maintained using corresponding a sin and cos function for each odd and even positions. Given enough GPU computational power Transformers can theoretically save infinitely long correspondence relationship among the word vectors, while in case of RNN/LSTM longer sequences tend to loss its initial information. Along with that as the data is not processed one by one it is highly efficient in case of parallelization. Figure 5 gives a complete overview.

## 2.5 Vision Transformer (ViT)

ViT ([Dosovitskiy et al., 2020](#)) is an approach to use transformer architecture for image classification without any convolutions. This approach surpasses all previous state of the art models for image classification. They borrow the architecture from original transformer encoder but their inputs embeddings are different. Initially, they break down the actual image into  $16 \times 16$  patches and flatten them one by one. After flattening, they tag a positional encoding so that the model can identify the sequence.

## 3 Related Works

The revolution of CNN lead to a lot of success in image processing domain. There are multiple significant models producing state of the art results in various multiclass classification problem. There are works of leveraging X-ray images to predict certain diseases. Like ([Ausawalaithong et al., 2018](#)) used DenseNet-121 ([Huang et al., 2017](#)) to predict lung cancer from X-ray images of chest. Even in recent years ([Narin et al., 2020](#)) worked on predicting Coronavirus using X-ray images using ResNet architecture. But in case of report generating a combined model is needed

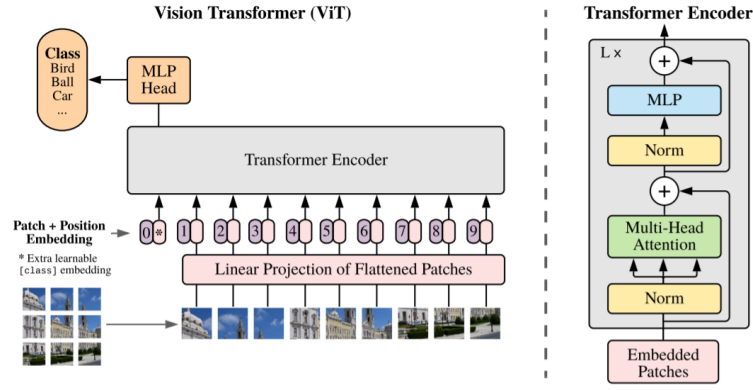


Figure 4: Vision Transformer

that can simultaneously work image data as well as process textual information. Recent works can be found in generating text from regular images in tasks like image captioning. But in the medical domain in case of report generation not much work has been done. Because unlike image captioning the required text is quite long and thus the sequential informational dependency is increased. (Dong et al., 2017) combined CNN and RNN firstly to extract exact diseases using X-ray images and then to describe the disease to generate corresponding report. Similar work has been done by (Shin et al., 2016) where they tried to classify/determine the disease from image and then produce a description containing the context of it. They also used Convolutional Neural Networks (CNNs) to process the images and RNN-LSTM based model for generating the text. One of the pioneering work was done by (Wang et al., 2018) They proposed a novel Text-Image Embedding network (TiNet) introducing multi-level attention based on CNN-RNN architecture. They integrated an end-to-end model leveraging both chest X-rays and corresponding reports to classify disease and then configured it to produce a primary report along with the classification given any new X-ray images. (Jing et al., 2017) proposed a multi-task learning framework for effectively predicting the keywords regarding the critical information in the findings identified as MIT (Medical Text Indexer) and generating a text description for which they built hierarchical LSTM using co-attention network. (Liu et al., 2019) also worked on generating accurate chest X-ray report. Their model first processed the image through Image Encoder and then to Sentence Decoder. The main idea was to build a domain aware model, that can at first detect the relevant topics and areas, and then can generate text on the basis. They used reward based system to keep the clinical coherent relevancy and intact the natural language generation process. In the paper of (Miura et al., 2020) they implemented a Meshed-Memory extended Transformer model which is a resemblance of CNN architecture encoded with a memory augmented attention process. The processed information from the images are then sent through a transformer architecture to decode into textual reports of the X-ray images.

## 4 Proposed Methodology

Figure 6 shows our proposed approach. We break that down to three basic steps.

### 4.1 Extracting and encoding image features

We break down the images into patches just like the original ViT (Dosovitskiy et al., 2020) and feed the embedding to the ViT encoder. We remove the MLP and Softmax layer. This approach results us with an encoder representation.

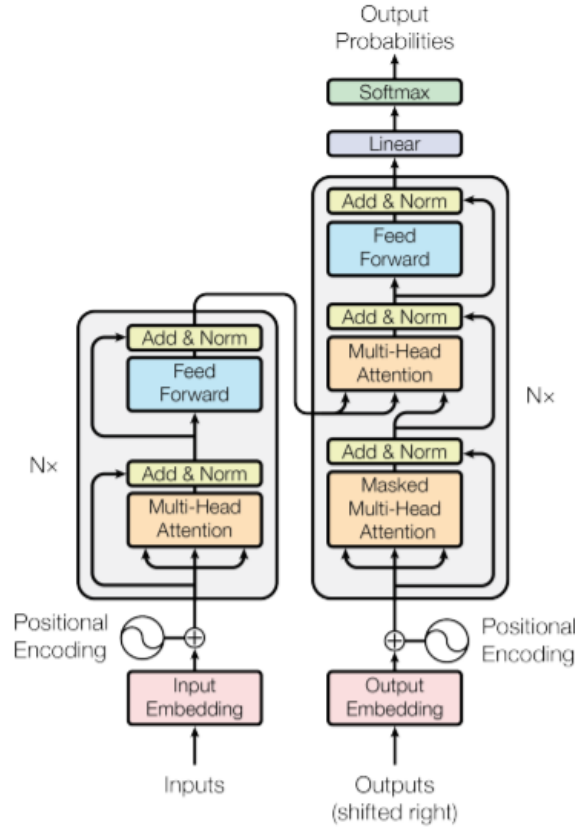


Figure 5: Transformer architecture

## 4.2 Decoding the encoded images with transformer decoder

We pass the encoded image to original transformer (Vaswani et al., 2017) decoder. This part of the architecture generates a readable and sensible text to generate the report.

## 4.3 Cross-Attention

We need to find out which part of the report attends to which part of the X-Ray image. We take the query and key from the vision transformer encoder and value from the report.

# 5 Experimental Analysis

## 5.1 Dataset

For our experiment, we used MIMIC-CXR (Johnson et al., 2019) dataset which is a publicly available database of chest radiographs with free-text reports. Since it is public, they collected and curated the data so that nobody can perform a linkage attack to trace the users. They did the annotations under expert supervision. The whole dataset contains more than 300,000 X-Ray and report pair which is by far the largest dataset for this task.

## 5.2 Experimental Setup

We did all our experiments using Google Colab which is a hosted Jupyter notebook service. We used it because Google Colab provides free GPU for 12 hours a day. In our experiments we used Numpy, Pandas, etc. for data processing and PyTorch for training and testing. PyTorch (Paszke et al., 2017) is an open source machine learning framework. For our experimentation, we used 30,000 text-image pair for train set, 3,000 for validation set and 3,000 for test set.

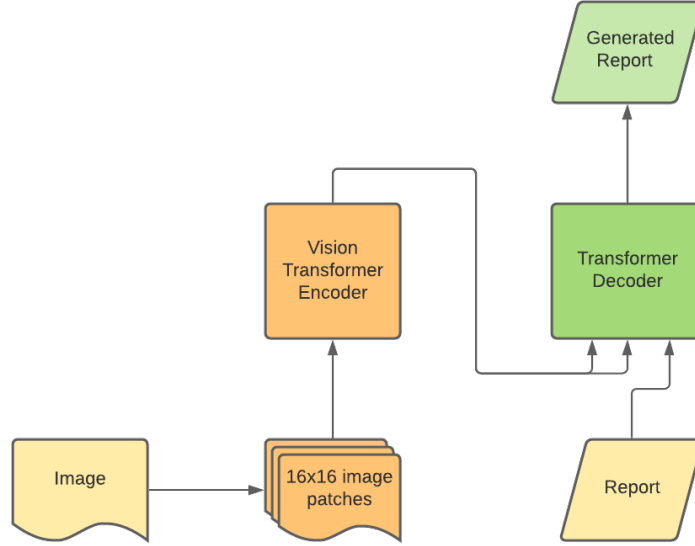


Figure 6: Full Transformer Architecture (Ours)

### 5.3 Result Analysis

Our approach works close to ResNet+Transformer in terms of Grammatical Correctness but works much worse than ResNet + Transformer in terms of Content Similarity since we trained with a small subset of the dataset. Table 1 shows a comparative analysis with the existing approaches.

Table 1: Comparative result analysis

Approach	Perplexity	BLEU
ResNet+Transformer	14.55	15.3
Full Transformer (Ours)	15.87	9.55

## 6 Conclusion and Future Plans

In our small scale experimentation, we got promising results but we are not satisfied with our results. In future, we plan to train and test our architecture with full extent of the dataset. We have further plans on data augmentation and adversarial fine-tuning for better results. Last but not least, we will keep looking for better and efficient approaches to solve this problem until we can find one.

## References

- W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn. 2018. [Automatic lung cancer prediction from chest x-ray images using the deep learning approach](#). In *2018 11th Biomedical Engineering International Conference (BMEiCON)*, pages 1–5.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Y. Dong, Y. Pan, J. Zhang, and W. Xu. 2017. [Learning to read chest x-ray images from 16000+ examples using cnn](#). In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 51–57.



- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.
- Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.
- Yasuhide Miura, Yuhao Zhang, Curtis P Langlotz, and Dan Jurafsky. 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*.
- Ali Narin, Ceren Kaya, and Ziyne Pamuk. 2020. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.
- Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058.