Radiology Report Generation using Full Transformer Architecture

Mohammad Sabik Irbaz

ID : 160041004
sabikirbaz@iut-dhaka.edu

Abir Azad

ID : 160041024
abirazad@iut-dhaka.edu

Supervisor

Sabbir Ahmed

Lecturer, Department of Computer Science and Engineering
sabbirahmed@iut-dhaka.edu

Islamic University of Technology

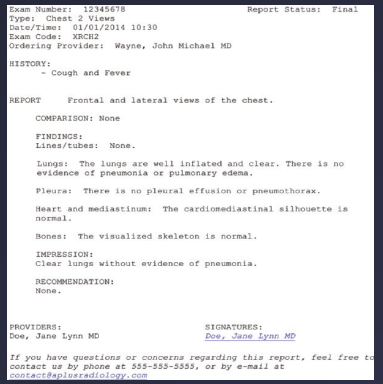
TABLE OF CONTENTS

01

Introduction

Introduction

- Image-to-text radiology report generation systems offer the potential to accelerate clinical processes by saving radiologists from the repetitive labor of drafting radiology reports and preventing medical errors.
- Our main focus of this research project is to find an efficient and scalable approach for report generation task.



Exam Number: 12345678 Report Status: Final
Type: Chest 2 Views
Date/Time: 01/01/2014 10:30
Exam Code: XRCH2
Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

RECOMMENDATION:
None.

PROVIDERS:
Doe, Jane Lynn MD

SIGNATURES:
Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

Motivation

- Automatic report generation from X-Ray images 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.

- Contributing to this research will help in decreasing the common medical errors by radiologists which sometimes cause major health issues.

Problem Statement

Given an X-Ray image
can the machine learn to understand and generate medical report from it?

02

Related Works

Existing approaches using MIMIC-CXR Dataset

MIMIC-CXR Dataset [Johnson et. al., 2019]

300,000+ images from over 60,000 patients

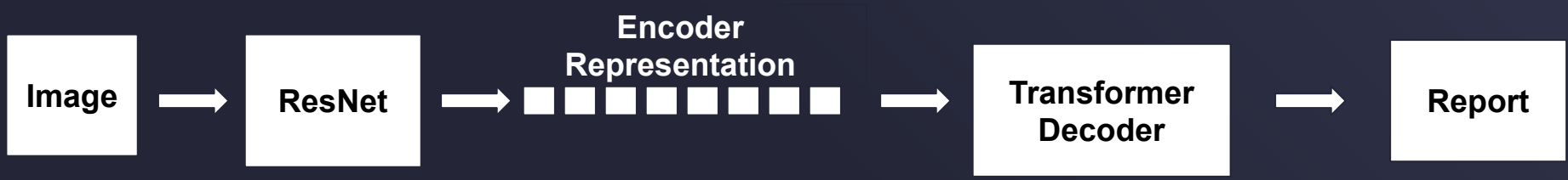
The large size of MIMIC-CXR allows for unprecedented studies in automated radiograph analysis.

X-Ray

Images	Reference
	PA and lateral views of the chest are obtained. There is mild atelectasis at the left lung base. 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 moderate pulmonary edema, but no pleural effusion or pneumothorax. Heart size is top-normal, stable. Mediastinal contours are within normal limits. Osseous structures are intact.

Report

ResNet + Transformer [Miura et al. 2020]



03

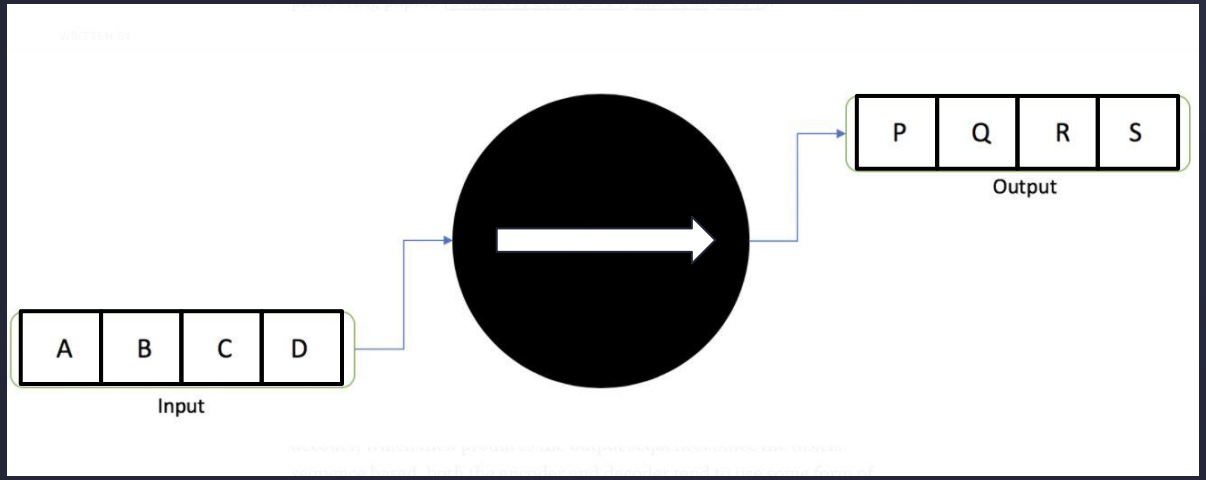
Background Study

Intuitions and analogies for our architecture.

Generic Translation Model

Sentence ⟷ **Sentence**

Sequence of words — **Sequence** of words

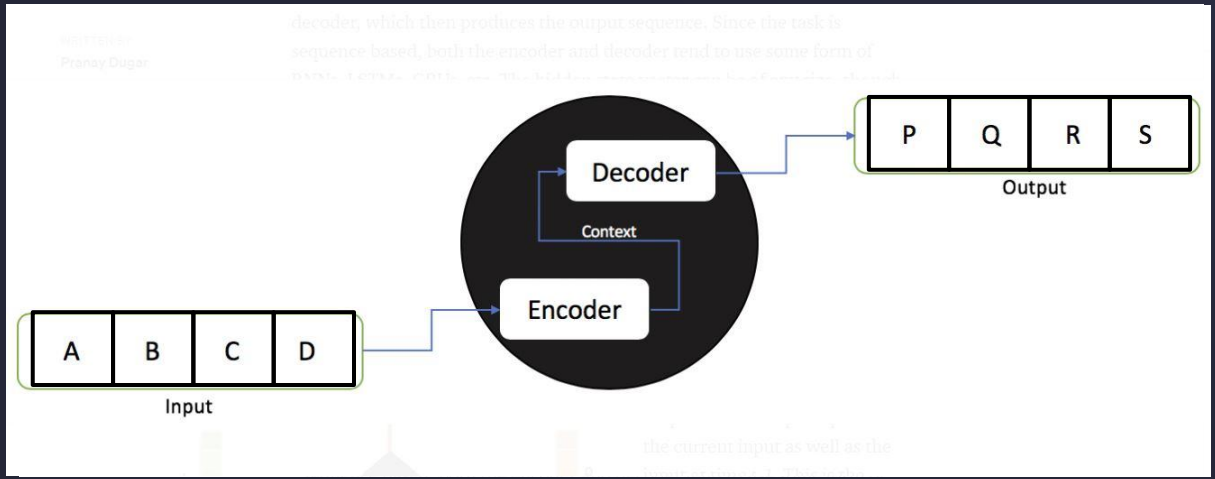


Generic Translation Model

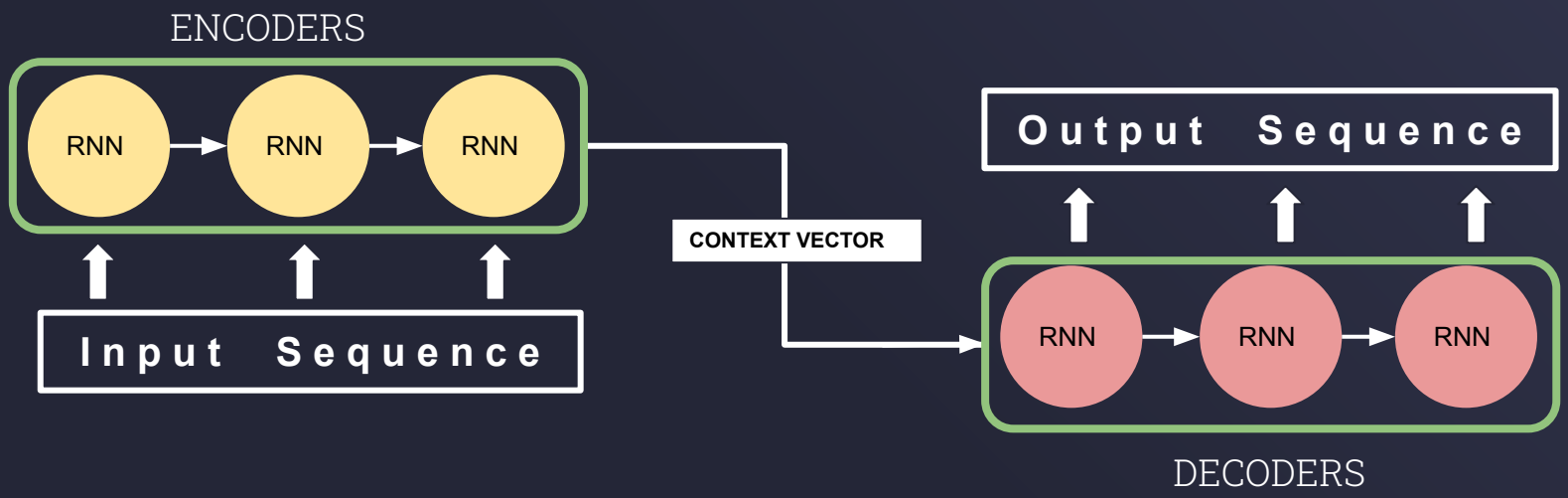
Sentence ⟷ **Sentence**

Sequence of words → **Sequence** of words

Seq2Seq Model



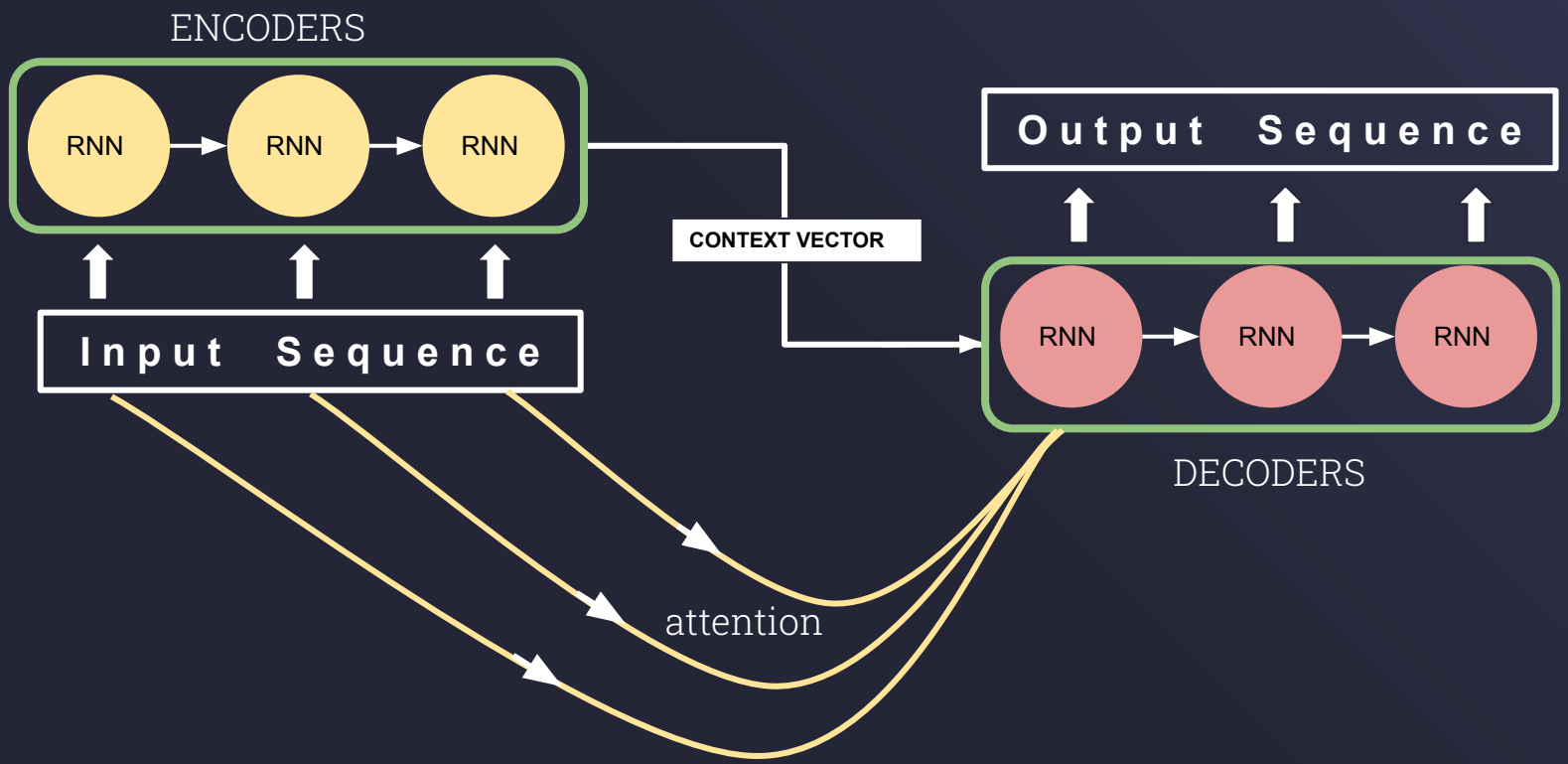
Encoder – Decoder [Sutskever et. al. 2014]



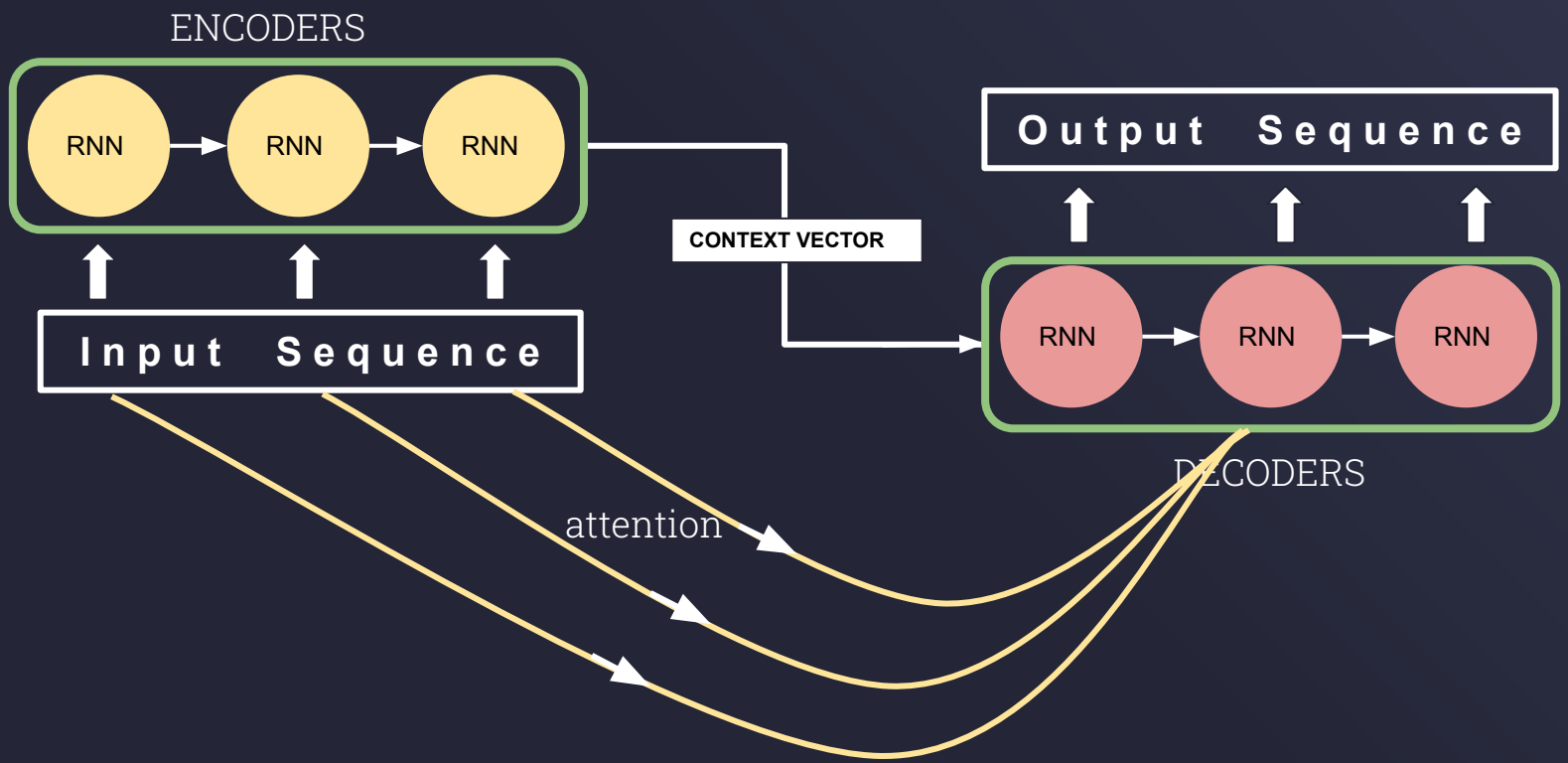
Drawbacks:

- The output sequence relies heavily on the context defined by the hidden state in the final output of the encoder
- In the case of long sequences, there is a high probability that the initial context has been lost by the end of the sequence.

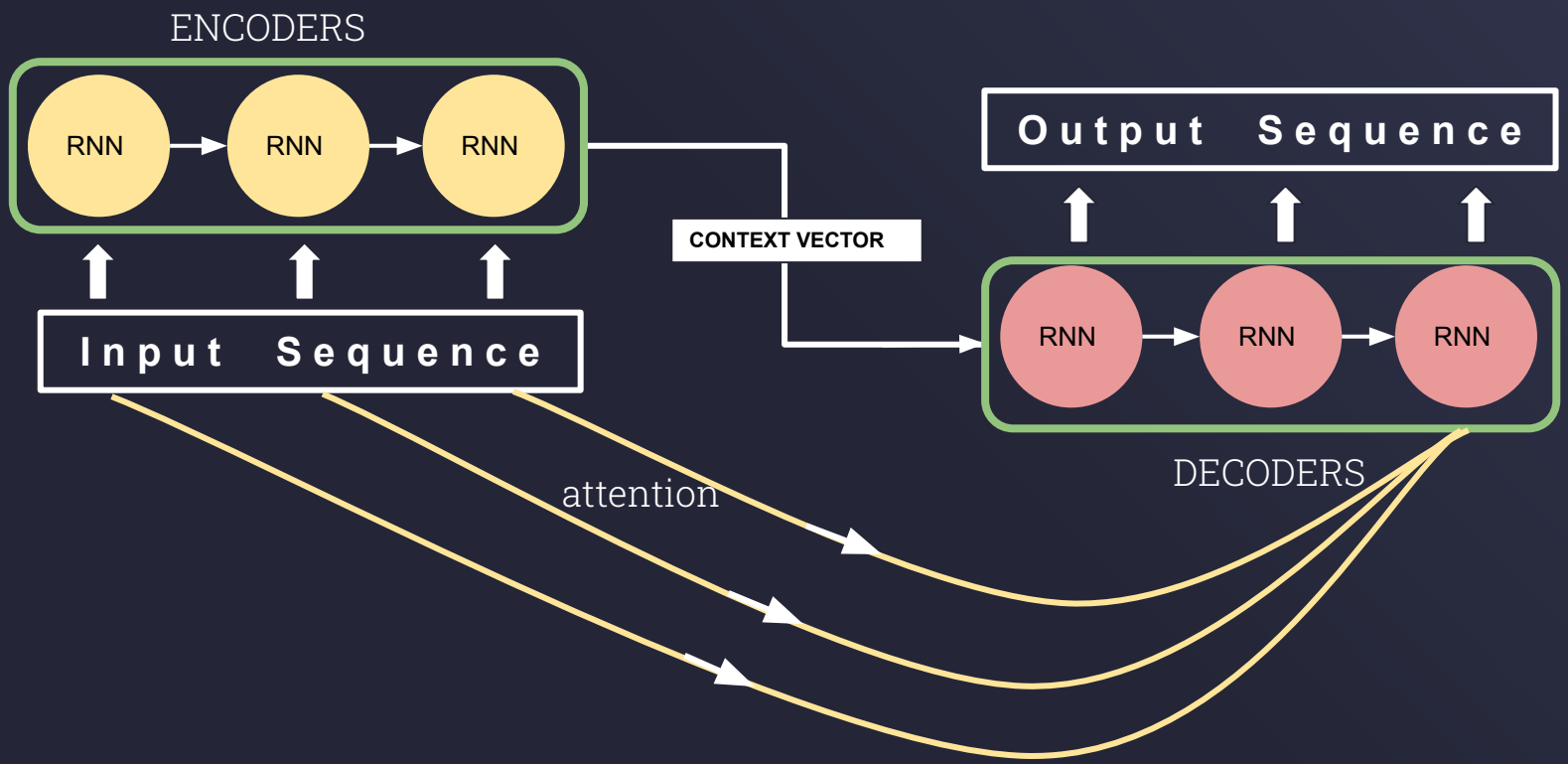
Encoder – Decoder with Attention [Bahdanau et. al., 2015]



Encoder – Decoder with Attention [Bahdanau et. al., 2015]



Encoder – Decoder with Attention [Bahdanau et. al., 2015]



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding

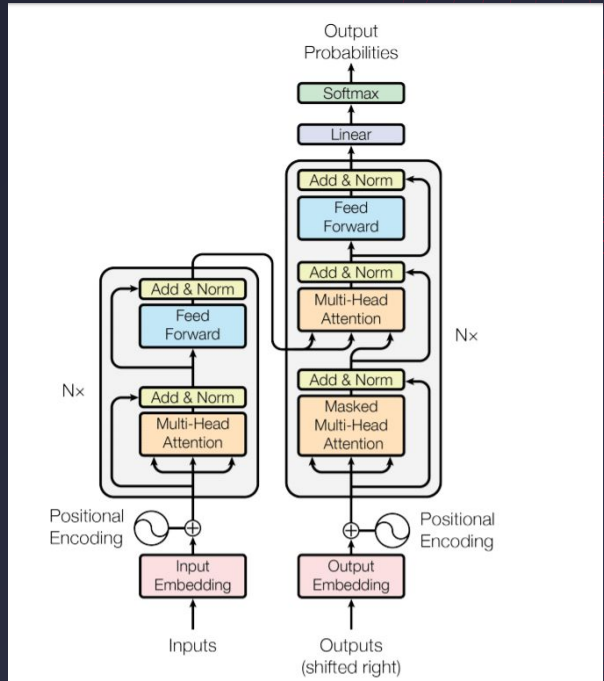


Figure 1: The Transformer - model architecture.

Transformers [Vaswani et. al., 2017]

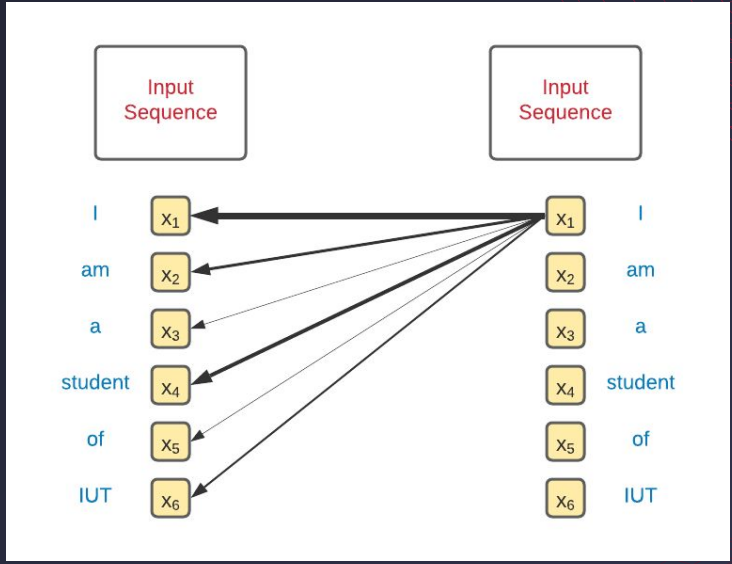
Proposed in the paper “Attention is all you need”

- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding

Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

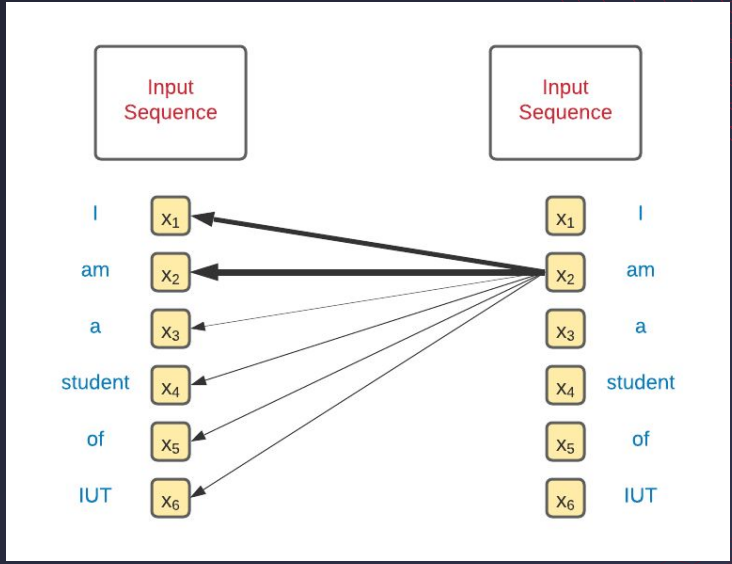
- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

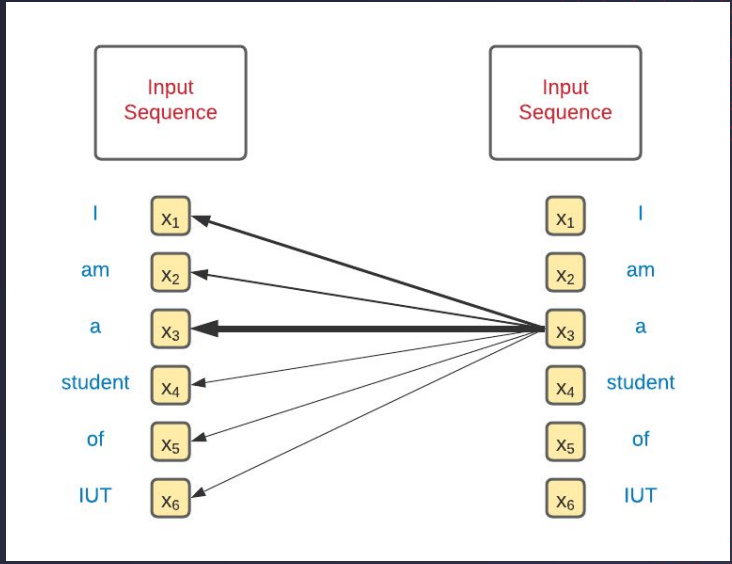
- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

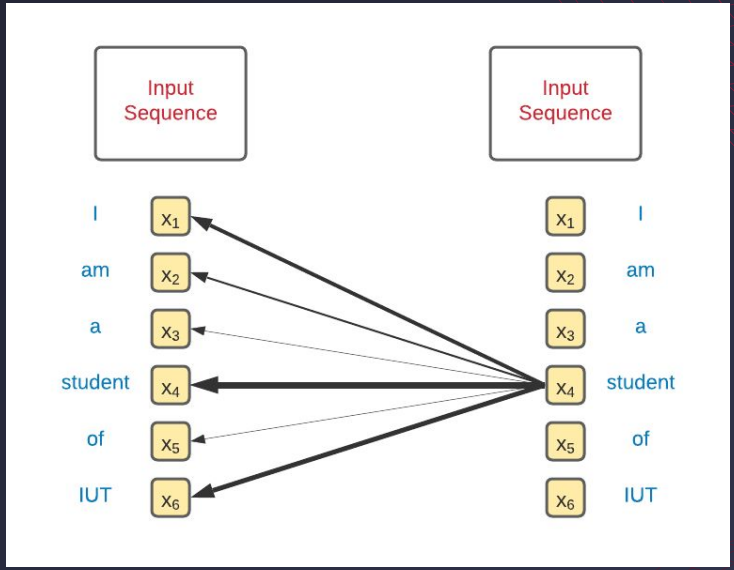
- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

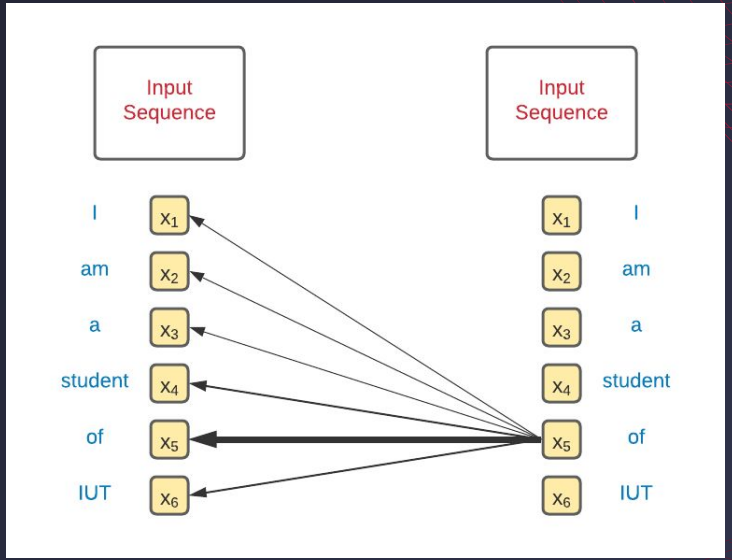
- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

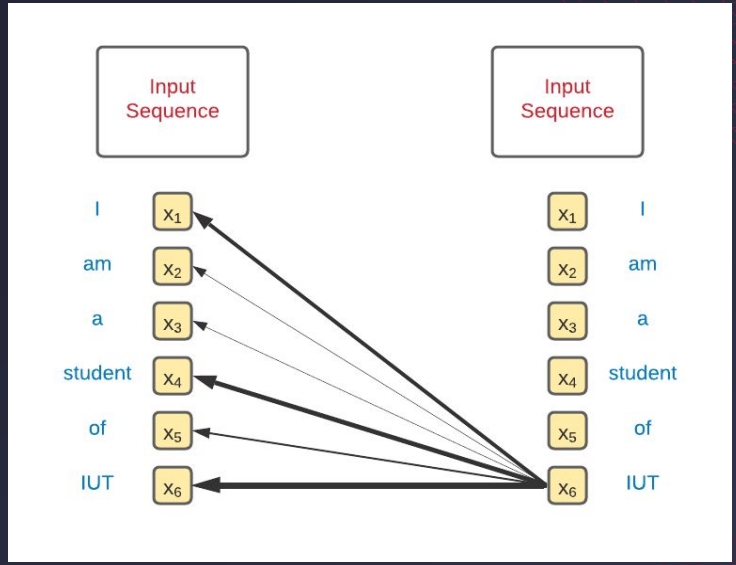
- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding

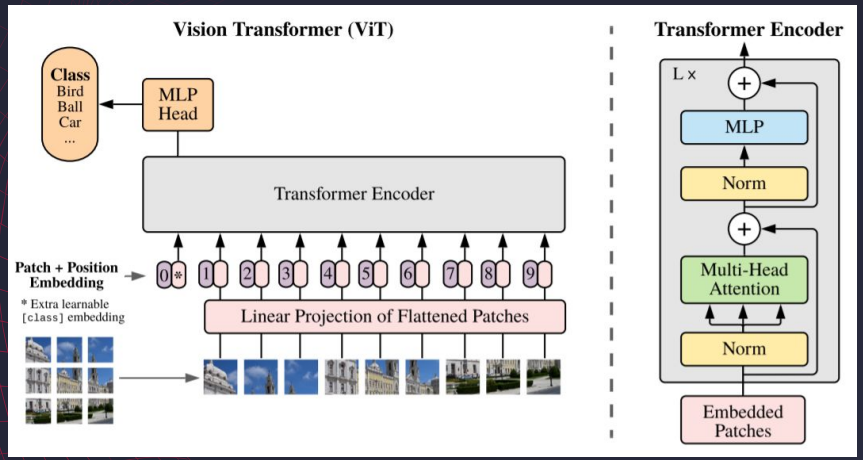


Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

- ➔ Attention
- ➔ Multi-head attention (self attention)
- ➔ Positional Encoding

Vision Transformer [Dosovitskiy et. al., 2020]



- **Patch Embedding :** An image in broken down into patches of 16 x 16. Then the patches are flattened and later given a sequence (for positional encoding).

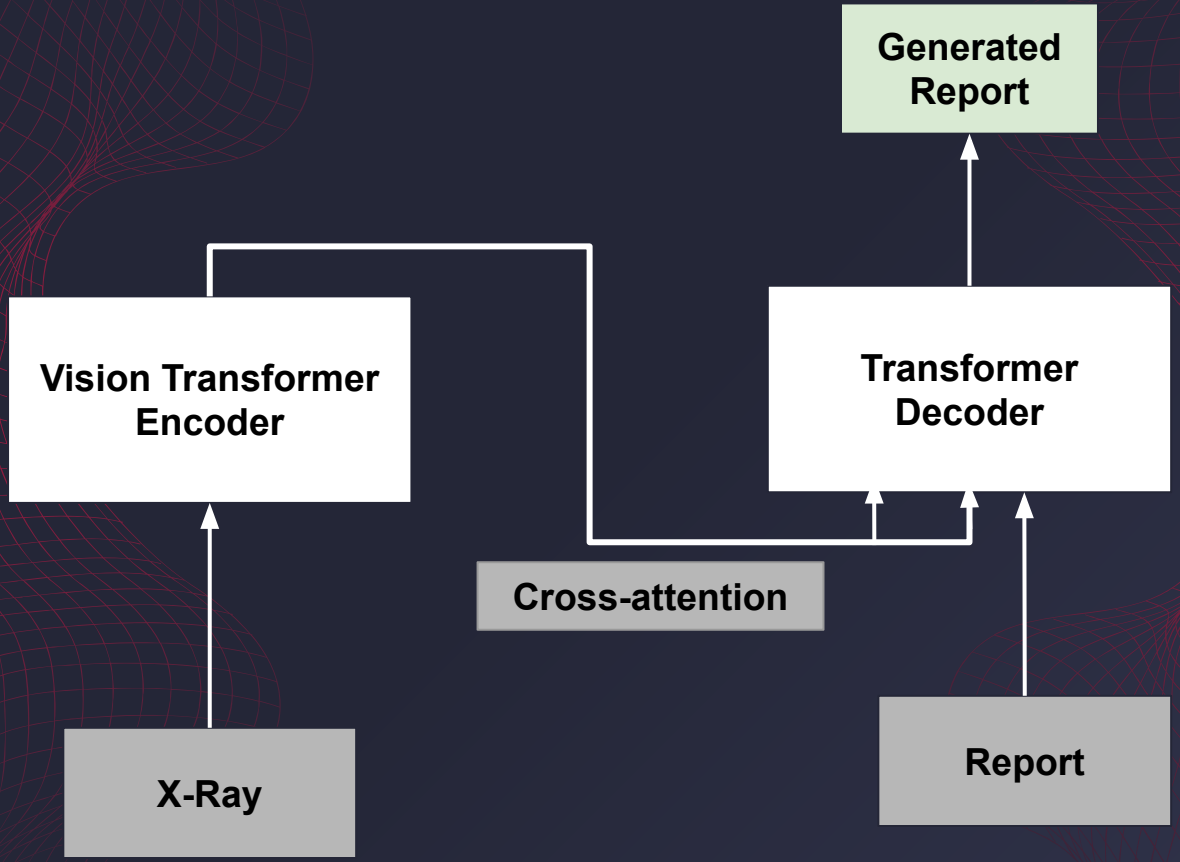
- The rest of the architecture is just like transformer encoder.

04

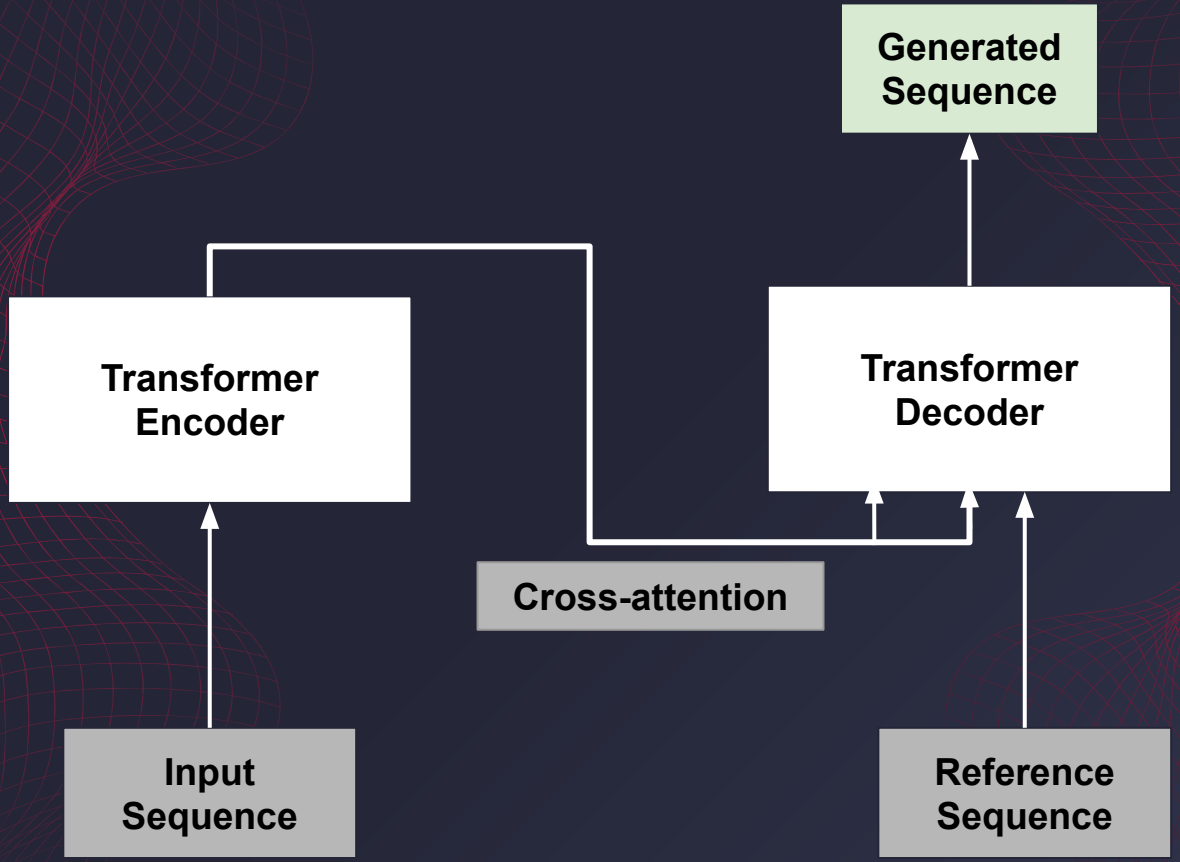
Proposed Methodology

New ideas, propositions and experimental results

Proposed Approach



Proposed Approach



Transformers [Vaswani et. al., 2017]

Proposed in the paper “Attention is all you need”

- ➔ Multihead Attention
- ➔ Positional Encoding
- ➔ Feed Forwarding Layer

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 is an open source machine learning framework.
- From the MIMIC-CXR dataset, we took 30,000 X-Ray and reports as train set, 3,000 as validation set and 3,000 as test set. We tuned the model on validation set and finally tested on the test set.

Evaluation Metrics

Natural Language Generation (NLG) tasks are evaluated by Perplexity and Content Similarity (BLEU)

BLEU (Bi-Lingual Evaluation Understudy) : [Papineni et. al., 2002]

We used tetra-gram (sequence of 4 consecutive words) BLEU score for this evaluation. To measure this, we check how many times all unique overlapping consecutive sequence of 4 words of the generated sentence appeared in target laymen sentence and divide it with number of tetragrams.

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

Perplexity : Checks the grammatical correctness of the generated texts.

$$\text{perplexity} = \prod_{t=1}^{T} \left(\frac{1}{P_{\text{LM}}(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})} \right)^{1/T}$$

Result Analysis

Grammatical Correctness

Approaches	Perplexity
Resnet+Transformer	14.55
(Ours)	**15.87**

Content Similarity

Approaches	BLEU
Resnet+Transformer	15.3
(Ours)	**9.55**

- ➔ Our approach works close to ResNet+Transformer in terms of Grammatical Correctness.
- ➔ Our approach works much worse than ResNet + Transformer in terms of Content Similarity.

05

Implementation Walkthrough

Conclusion and Future Plans

- We proposed a new architecture for report generation which doesn't depend on convolutions anymore.
- We are planning to train and test our approach with full dataset.
- We are also planning to come up with an approach that can take image features from multiple images for one report.
- We have further plans on data augmentation and adversarial fine-tuning for better results.
- Further human evaluation after training and testing due to the deficiency of proper evaluation metrics.