

# **Neural Models for Generating Clinically Accurate Chest X-Ray Reports**

**André Loras Leite**

Thesis to obtain the Master of Science Degree in

## **Information Systems and Computer Engineering**

Supervisors: Prof. Bruno Emanuel Da Graça Martins  
Prof. Arlindo Manuel Limede de Oliveira

### **Examination Committee**

Chairperson: Prof. Alberto Manuel Rodrigues da Silva  
Supervisor: Prof. Bruno Emanuel Da Graça Martins  
Member of the Committee: Prof. David Martins de Matos

**October 2022**



# Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.



# Acknowledgments

I want to share my gratitude to both my dissertation advisors, Prof. Bruno Martins and Prof. Arlindo Oliveira, for all the guidance and shared knowledge.

I also want to give recognition to INESC-ID, since without their infrastructure this dissertation would have not been so successful.

I would like to thank my friends, who have been there for me in the worst and best moments of this journey. I want to thank all of them for their constant support and friendship.

I would also like to show my deep appreciation for everything that my parents and brother have done to support me. I would like to thank them for all the love and care they have shown me throughout my academic years. This is also their victory.

To all, I thank you.



# Abstract

Image captioning models have been increasing their performance comprehensively, having shown that artificial intelligence is capable of achieving successful results in computer vision tasks. However, there are still some tasks within the range of image captioning that need more focus, including the automatic clinical report generation. The automatic generation of radiology reports based on radiology images has gathered an increasing amount of focus in the last few years. This is supported by the repetitive and exhaustive work that these clinical reports demand. Artificial neural networks that address this task have been changing over the years, starting as convolutional neural networks, changing over to transformer-based models. However, these existing methodologies focus more on one of two important aspects, that being the fluency and human-readability capacity of the generated text, over the clinical efficiency of the model. Consequently, in this dissertation we propose a model capable of achieving competitive results regarding the human readability of the reports, as well as improving clinical efficiency. We propose to adapt the MedCLIP model to have an image-text encoder capable of concatenating both image and text. We further propose that this model works with the assistance of an Information Retrieval mechanism (i.e. FAISS), to retrieve reports that are resultant of a similarity evaluation done on an input x-ray, obtaining the closest reports. On the MIMIC-CXR dataset, our model has improved on both natural language processing metrics and clinical efficiency, over well established models. Finally, we further show that our model can lead to more human-readable reports, while keeping clinical actuality, over most state-of-the-art models.

## Keywords

Artificial Neural Network; Convolutional Neural Networks; Transformers; Radiology Images; Natural Language Processing; Computer Vision; Image Captioning; Information Retrieval Mechanism.





# Resumo

Os modelos de legendagem de imagens têm vindo a melhorar o seu desempenho, tendo demonstrado que a inteligência artificial pode alcançar resultados notórios em tarefas de visão computacional. No entanto, ainda existem algumas tarefas semelhantes que necessitam de igual atenção, incluindo a geração automática de relatórios clínicos. A geração automática destes relatórios com base em imagens de radiologia tem vindo a reunir um número crescente de atenção. Isto é apoiado pelo trabalho repetitivo e exaustivo que estes relatórios clínicos exigem. As redes artificiais que abordam esta tarefa têm vindo a mudar, começando como redes convolucionais, mudando para modelos baseados em transformers. Contudo, estas metodologias existentes centram-se mais num de dois aspectos importantes, que é a fluência e a capacidade de leitura humana do texto gerado, sobre a eficiência clínica do modelo. Consequentemente, nesta dissertação propomos um modelo capaz de alcançar resultados competitivos no que diz respeito à legibilidade humana dos relatórios, bem como de melhorar a eficiência clínica. Propomos também adaptar o modelo MedCLIP de forma a ter um encoder de imagem-texto capaz de concatenar tanto a imagem como o texto. Propomos ainda que este funcione em par com um mecanismo de recuperação de informação (e.g., FAISS), para recuperar relatórios que resultem de uma avaliação de similaridade feita com base num raio-x, obtendo os relatórios mais próximos. No conjunto de dados do MIMIC-CXR, o nosso modelo melhorou tanto a métrica de processamento da linguagem natural como a eficiência clínica, em relação a modelos bem estabelecidos na área. Finalmente, mostramos ainda que o nosso modelo pode gerar relatórios mais legíveis, mantendo a factualidade clínica.

## Palavras Chave

Redes Neurais; Redes Convolucionais; Transformers; Imagens de Radiologia; Processamento de Linguagem Natural; Visão Computacional; Image Captioning; Mecanismo de Recuperação de Informação.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Context and Motivation . . . . .	3
1.2	Objectives . . . . .	4
1.3	Summary of Contributions . . . . .	4
1.4	Thesis Outline . . . . .	5
<b>2</b>	<b>Fundamental Concepts</b>	<b>7</b>
2.1	The Perceptron . . . . .	9
2.2	Artificial Neural Networks . . . . .	10
2.3	Recurrent Neural Networks . . . . .	12
2.4	Convolutional Neural Networks . . . . .	13
2.5	Transformers . . . . .	14
<b>3</b>	<b>Related Work</b>	<b>19</b>
3.1	Report Generation Based on Transformers . . . . .	21
3.1.1	Generating Radiology Reports Via Memory-Driven Transformer . . . . .	21
3.1.2	Progressive Transformer-Based Generation of Radiology Reports . . . . .	23
3.1.3	Learning to Generate Clinically Coherent Chest X-ray Reports . . . . .	24
3.1.4	Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation . . . . .	24
3.1.5	Automated Generation of Accurate & Fluent Medical X-ray Reports . . . . .	25
3.1.6	Weakly Supervised Contrastive Learning for Chest X-ray Report Generation . . . . .	26
3.1.7	Aligntransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation . . . . .	26
3.2	General Image Captioning . . . . .	27
3.2.1	How Much Can CLIP Benefit Vision-and-Language Tasks? . . . . .	27
3.2.2	ClipCap: CLIP Prefix for Image Captioning . . . . .	28
3.2.3	CPTR . . . . .	29
3.2.4	VL-T5 and VL-BART . . . . .	30

3.3	Retrieval Mechanism in Image Captioning Tasks . . . . .	31
3.3.1	Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model . . . . .	32
3.3.2	Retrieval-Augmented Image Captioning . . . . .	33
3.3.3	Retrieval-Augmented Transformer for Image Captioning . . . . .	33
3.3.4	Memory-Augmented Image Captioning . . . . .	34
3.4	Overview . . . . .	34
<b>4</b>	<b>Methodology</b>	<b>37</b>
4.1	Encoder-Decoder Architecture . . . . .	39
4.1.1	Encoder . . . . .	40
4.1.2	Decoder . . . . .	41
4.2	Retrieval Mechanism . . . . .	41
4.3	CLIP V&T . . . . .	42
<b>5</b>	<b>Experimental Evaluation</b>	<b>43</b>
5.1	Experimental Setup . . . . .	45
5.1.1	Dataset . . . . .	45
5.1.2	Evaluation Metrics . . . . .	45
5.1.3	Hyperparameters Fine-tuning . . . . .	45
5.1.4	Evaluation Methodology and Training . . . . .	46
5.2	Experimental Results . . . . .	47
5.2.1	Report Generation Models Performance . . . . .	47
5.2.2	Clinical Efficiency Evaluation . . . . .	48
5.2.3	Retrieval Mechanism Evaluation . . . . .	49
5.2.4	Report Generation Study . . . . .	51
5.2.5	Contrastive Attention Study . . . . .	52
5.2.6	Overview . . . . .	53
<b>6</b>	<b>Conclusion</b>	<b>55</b>
6.1	Conclusions . . . . .	57
6.2	System Limitations and Future Work . . . . .	57
	<b>Bibliography</b>	<b>59</b>

# List of Figures

1.1	Example of radiology images from the dataset MIMIC-CXR <a href="#">Johnson et al. (2019)</a> . . . . .	3
2.1	Perceptron Model. . . . .	9
2.2	Basic structure of an Artificial Neural Network (ANN). . . . .	10
2.3	Basic structure of a Recurrent Neural Network (RNN). . . . .	12
2.4	Typical Convolutional Neural Network architecture. . . . .	13
2.5	Transformer model following the specifications of <a href="#">Vaswani et al. (2017)</a> . . . . .	15
2.6	Single attention and Multi-head attention, according to <a href="#">Vaswani et al. (2017)</a> . . . . .	16
3.1	Architecture of a Memory-driven Transformer RM+MCLN ( <a href="#">Chen et al., 2020</a> ). . . . .	22
3.2	Framework of the $M^2$ Tr. Progressive ( <a href="#">Nooralahzadeh et al., 2021</a> ). . . . .	23
3.3	Overview of ClipClap transformer-based architecture. . . . .	29
3.4	Architecture of CPTR ( <a href="#">Liu et al., 2021c</a> ). . . . .	30
3.5	Architecture of VL-T5 and VL-BART ( <a href="#">Cho et al., 2021</a> ) for visual grounding task. . . . .	31
4.1	Architecture of the CLIP VT Model with retrieval mechanism. . . . .	39
5.1	Retrieval process for $k = 3$ nearest neighbors and respective reports. . . . .	51
5.2	Three randomly chosen results of the CLIP VT w/ Retrieval model, showing the practical results on the generation of new clinical reports. . . . .	52
5.3	Contrastive attention study done on two randomly chosen x-rays from the MIMIC-CXR dataset. . . . .	53



# List of Tables

3.1	Summary of the models presented in the related work section. . . . .	34
5.1	Beam Search decoding method evaluation on the CLIP vision and text model. . . . .	45
5.2	NLP performance results containing the MIMIC-CXR and IU-Xray datasets. . . . .	48
5.3	Clinical accuracy results on some of the Transformer Models. . . . .	49
5.4	Retrieval mechanism evaluation (of CLIP VT w/ Retrieval) on the capacity of retrieving reports with close meaning to the original report, regarding image-to-image similarity, with k neighbors variation. . . . .	50
5.5	Retrieval mechanism evaluation (of CLIP VT w/ Retrieval) on image-to-image similarity compared to image-to-text similarity. . . . .	50
5.6	Comparison between some results of the retrieval mechanism. . . . .	50





# 1

## Introduction

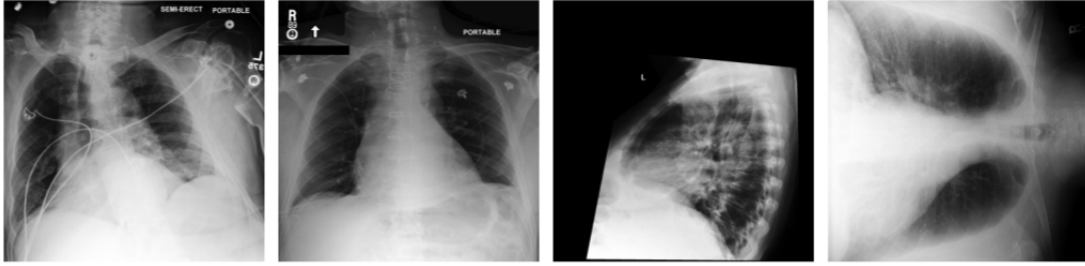
### Contents

---

1.1	Context and Motivation . . . . .	3
1.2	Objectives . . . . .	4
1.3	Summary of Contributions . . . . .	4
1.4	Thesis Outline . . . . .	5

---





**Figure 1.1:** Example of radiology images from the dataset MIMIC-CXR [Johnson et al. \(2019\)](#).

In this chapter, we introduce the context of the problem we propose to mitigate or solve with the methodologies presented in this MSc. dissertation, followed by the motivation to do so. We also present the objectives achieved considering the duration of the present work. Following this, is a summary of contributions, where we indicate what were the main additions to this scientific area. Conclusively, we present the organization of the document.

## 1.1 Context and Motivation

The automatic generation of radiology reports, given medical x-rays as inputs, has significant potential to facilitate administrative operations and improve clinical patient care. Several previous studies have focused on this problem, employing methods from computer vision and natural language generation to produce readable reports. Typical solutions are based on encoder-decoder neural network architectures, in which an encoder component produces intermediate representations from the input visual contents, and then a decoder component generates the target report token-by-token. Although the aforementioned typical approaches have achieved interesting experimental results, they often fail to account for the particular nuances of the radiology domain and, in particular, the critical importance of clinical accuracy in the resulting reports. In the context of my M.Sc. dissertation, we have explored neural models for chest X-ray report generation, extending previous methods in several directions, where we finally propose a model capable to generate competitive results.

Specifically, we propose (a) the use of Transformer sequence-to-sequence models similar to those used in other image captioning tasks ([Vaswani et al., 2017](#); [Liu et al., 2021c](#); [Cho et al., 2021](#); [Nooralahzadeh et al., 2021](#); [Shen et al., 2021](#); [Mokady et al., 2021](#); [Cornia et al., 2020](#); [Endo et al., 2021](#)), (b) the use of policy gradient methods to train the models using clinical coherence/factuality as a reward function ([Irvin et al., 2019](#); [McDermott et al., 2020](#); [Smit et al., 2020](#); [Ippolito et al., 2019](#)), (c) using information from similar training instances to guide the report generation process ([Liu et al., 2021b](#); [You et al., 2021](#); [Lovelace and Mortazavi, 2020](#); [Yan et al., 2021](#); [submission, 2022](#); [Wang et al., 2020](#); [Xu et al., 2021](#)),

or (d) using alternative decoding methods that reorder a set of diverse alternative generations according to clinical coherence/factuality (Zarrieß et al., 2021).

Experiments will be performed on one of the most well-known datasets in the area, specifically the MIMIC-CXR dataset (Johnson et al., 2019). Quality will be assessed in terms of text generation metrics such as BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2015), as well as in terms of clinical coherence/factuality using precision, recall, and the F1 score.

## 1.2 Objectives

The radiology report generation task is continuously being improved with some of the most promising methods of deep learning. However, in this dissertation, we proposed the use of Transformer architecture as the main tool to provide the mechanisms to enhance the task of generating text, based on the analysis of a radiology report. Furthermore, in this dissertation, we study the extent of improvements that CLIP (Radford et al., 2021a) is capable of presenting while considering other state-of-the-art models as the main comparison while using an information retrieval mechanism.

In detail, some of the more specific objectives are: (1) Explore the use of Transformer models against previous approaches such as Convolutional Neural Networks and LSTMs on the task of medical report generation; (2) Explore the capacity of the CLIP model to generate clinically accurate and semantically correct reports according to radiology images; (3) Combine both CLIP Text and CLIP Vision into a single encoder, adapt this encoder as the encoder of a firstly proposed baseline, and finally evaluate the performance; (4) Augment the encoder-decoder model architecture with a retrieval mechanism to guide the generation.

## 1.3 Summary of Contributions

We can summarize the main contributions of this MSc. dissertation on the following points:

- The proposal of a new encoder-decoder model, where the encoder is the combination of both CLIP Text and CLIP Vision encoders.
- Enhancement of the model with a retrieval mechanism, based on the FAISS similarity mechanism.
- The comparison of the new model against state-of-the-art models trained on medical data, improving clinical efficiency. The use of a combined encoder architecture showed benefits over the common vision encoder approach.
- Improvement of the human-readability of the newly generated reports.

- Improvement of the MEDClip model on the task of generating accurate medical reports by switching the decoder with a fine-tuned GPT-2 decoder instead of the BERT decoder. Also, the fine-tuning of the whole model by comparing it with other models.

## 1.4 Thesis Outline

This thesis is organized as follows: Chapter 1 presents the documents as well as the context that provides the motivation to employ such work, also providing crucial organizational notions over the present dissertation document. In chapter 2 we set forth a chapter solely dedicated to introducing the underlying foundations of this dissertation, providing theoretical notions over the most used methods on tasks like image captioning, which is very similar to the task that we report in this dissertation. In chapter 3 we provide a broad scale of works that support and try to solve problems very similar to the ones present in this thesis. In Chapter 5 we present in more detail the construction of the models that sustain this dissertation's objectives. In chapter 4 we provide a clear notion of the tools used to evaluate this proposal, followed by a spectrum of tests and their results. Finally, in Chapter 6 we establish the overall conclusions based on the whole dissertation followed by directions on future work and limitations of the presented approach.



# 2

## Fundamental Concepts

### Contents

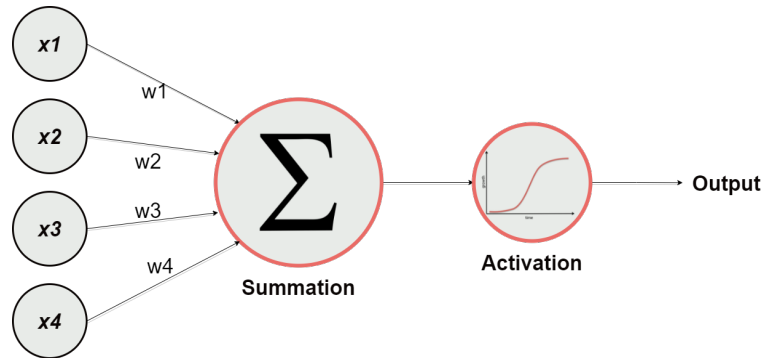
---

2.1 The Perceptron . . . . .	9
2.2 Artificial Neural Networks . . . . .	10
2.3 Recurrent Neural Networks . . . . .	12
2.4 Convolutional Neural Networks . . . . .	13
2.5 Transformers . . . . .	14

---







**Figure 2.1:** Perceptron Model.

This chapter focuses on the theoretical aspects that support this dissertation. We will provide background on concepts such as the Perceptron, Artificial Neural Networks, Recurrent Neural Networks, Convolutional Neural Networks, and finally Transformers.

## 2.1 The Perceptron

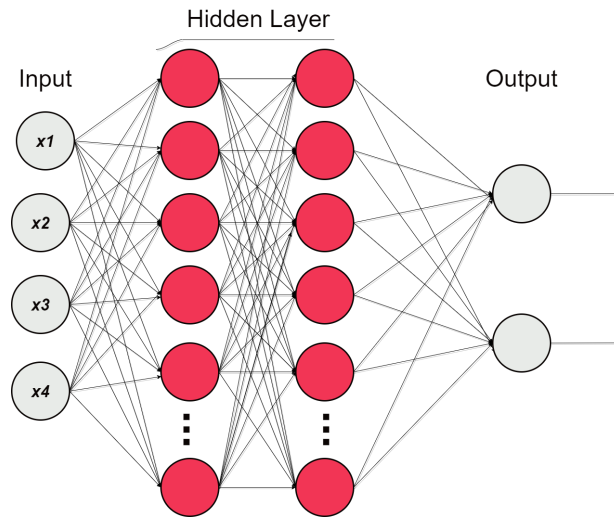
The human brain is inhabited by billions of neurons ( $\approx 86$  billion). The morphological description of a neuron concerns a nerve cell responsible for processing and transmitting electrical and chemical signals. These kinds of transfers are called synapses. Consequently, to provide such a foundation, neurons must be interconnected by some channel that can provide a gateway for passing signals. Dendrites are the “gateways” that allow the neurons to communicate.

Taking inspiration from biological neurons, the Perceptron was introduced by Frank Rosenblatt in 1957 (Rosenblatt, 1958), as a model and a learning rule based on the original McCulloch-Pitts neuron. A perceptron can be seen as an algorithm for supervised learning over binary classification tasks. It enables neurons to learn elements in a training set, one at a time. Individual perceptrons can only deal with linearly definable classification tasks, where the goal is to find a linear function involving a weight vector and a bias factor. Multilayer perceptrons, or feedforward neural networks, have numerous perceptrons organized into layers, thus having greater processing power, and leveraging more challenging classification tasks.

In order to enable a distinction between two linearly separable classes, 1 and 0, the perceptron acquires knowledge from the weights with respect to the input, in order to draw a linear decision boundary.

The original learning rule states that the algorithm would automatically learn the optimal weight coefficients. The input features are then multiplied with these weights to determine if a neuron should prompt action or stay idle.

More formally, the perceptron is a function that maps an input  $\mathbf{x}$  multiplied with the learned weight



**Figure 2.2:** Basic structure of an Artificial Neural Network (ANN).

coefficients, where  $f(x)$  can generate.

$$f(x) = \begin{cases} 1, & w \times x + b > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2.1)$$

$$\sum_{i=1}^m w_i \times x_i + b \quad (2.2)$$

According to Figure 2.1, we can state that the output will be a combination of the summation of the weights multiplied by the input defined by Equation 2.2, finally passed through an activation Equation 2.1 that will prompt if the perceptron should run or not. The activation function is defined by Equation 2.1, where  $\mathbf{w}$  is the vector of real-valued weights,  $\mathbf{b}$  stands for the bias (an element that adjusts the boundary away from origin without any dependence on the input value), and  $\mathbf{x}$  as the vector of input  $x$  values.

## 2.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are composed of individual perceptrons. There are usually three types of layers: input, hidden, and output layers.

As depicted in Figure 2.2 the hidden layer is located between the input and output layers. Similarly to what was explained in Section 2.1, the input to each node is factored by the weights, passing then through the activation function, and formulating an output. Hidden layers modulate the possible outcome

of a network, performing parallel transformations to some input, allowing for the network to break down into specific amendments of some data. Each layer, independently of its cardinality, will be responsible for outsourcing a very detailed result to the next one. Typically, as an activation function, it is used the ReLU, which stands for Rectified Linear Unit. Used as an improvement over Sigmoid or Tanh (hyperbolic tangent), the ReLU is a function that returns 0 when receiving negative inputs and returns the received value if it is positive. However, it is common also to use Softmax as an activation function, as it is capable to handle a multiple-class distribution. Thus, having more capacity on dealing with more sophisticated models, such as Transformers (as we will later discuss). Softmax, as is depicted in Equation 2.3, the output probabilities will be correlated, meaning that when summed, the total will always be 1, contrarily to the Sigmoid activation function, as it looks separately to each output value.

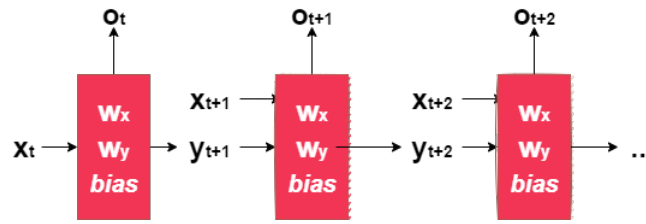
$$\sigma(y_i) = \left( \frac{e^{y_i}}{\sum_j e^{y_j}} \right) \quad j = 1, \dots, n \quad (2.3)$$

We cannot be certain of the hidden layer's ability to generate a good transformation, as this is completely dependent on the weights associated and the activation function in use (ReLU is a very commonly used activation function).

These neural networks learn on the basis that each node is present with a linear regression model, where we have the same inputs seen in the perceptron 2.1. However, to train this kind of network, the back-propagation algorithm is used. This algorithm is the foundation of the learning method in neural networks, and it is simply the fine-tuning of the weights and the bias factor, all considering the loss rate in each training epoch. As depicted in Figure 2.2, the leftmost layer considers the inputs for the network, which take the bias term before entering the middle layer, being the hidden layer. Finally, we get the output, which is actually the activation value, representing the actual model's decision. In backward propagation, the first step is called forward propagate, making information flow from one layer to the next, where first, we must pass the weighted sum of the inputs using Equation 2.1, and second, get the activation value from the activation function upon the weighted sum. However, this is not enough to get a correct prediction, considering the fact that the weights are static. To circumvent that, by adding a step of back-propagation, we can use the loss and feed it backwards, in order to fine-tune these weights. This happens as partial derivatives of the activation functions. To better suit these weights, gradient descent is often used.

$$a_{n+1} = a_n - \gamma \Delta f(a_n) \quad (2.4)$$

Gradient descent is described in Equation 2.4, where  $a_{n+1}$  is the next position, and consequently



**Figure 2.3:** Basic structure of a Recurrent Neural Network (RNN).

$a_n$  is the current position. As gradient descent minimizes a given cost function, this step is defined with a subtraction of the current position  $a$  with a waiting factor and the gradient ( $\gamma$  and  $\Delta f(a_n)$ ), giving the step for the more elevated descent. This steepest evolution will lead the gradient descent to generate a path leading to the cost function with the least value possible. Gradient descent has many derived methods like Batch Gradient Descent (BGD) or Stochastic Gradient Descent (SGD). In the BGD the error is retrieved for each data sample in the dataset, being updated when finally every error associated with each batch has been retrieved, happening in every epoch. The SGD provides an update of these parameters singularly, meaning that for each data batch, the SGD will update the parameters.

## 2.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are more suitable for working with time series data, or data that has a sequential structure. This leads to the fact that RNNs are models that retain memory from past states. The most common architecture that defines an RNN is shown in Figure 2.3:

Starting with an input state  $x_t$ , when going through the network, this input will produce a new state  $y_t$  that will now store a vector with the values computed in the hidden layer. This generated output from the first computation ( $y_t$ ), will now be fed again to the network, generating  $y_{t+1}$ . Due to this very specific architecture, the RNN falls short in many cases, such as the more steps the network takes, the less it actually learns. This is a consequence of getting the gradient as low as almost zero, disabling the learning curve for new weights. In order to overcome the limitations of the classic RNN, some new formulations have been proposed, such as Long Short-Term Memory (LSTM) or Gate Recurrent Units (GRUs). In contrast to the common feedforward neural network, LSTMs are designed with a feedback component, which is capable of circumventing the vanishing gradient problem, raised in RNNs. Thus, providing a method for solving the long-term dependency, common of RNNs. As they retain information from previous data, LSTMs are able to process large sequences of data without having to treat each point in the sequence separately. Also being considered a variation of LSTMs, where both designs are similar, GRUs are also an improvement over RNNs to solve the vanishing gradient problem. The difference between GRUs and LSTMs is in the process created to tackle this problem. Implemented in

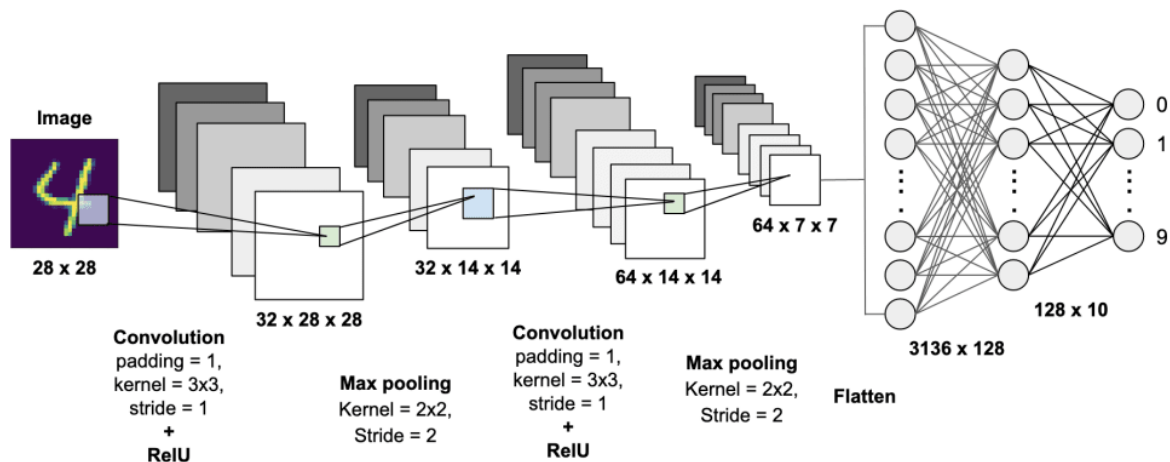


Figure 2.4: Typical Convolutional Neural Network architecture.

GRUs are an update gate and reset gate which is basically two vectors deciding what data goes to the output.

## 2.4 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have been massively in use for Computer Vision tasks. For some typical problems regarding Computer Vision, the architecture presented by this network really enhances the results. In fulfillment of this dissertation, CNNs have been one of the most desired methodologies to solve the resolution of radiology report generation. Thus, is it imperative to sustain a more concrete notion of what makes this deep learning algorithm so desirable. The common network is composed of simple layers, such as the input or hidden layer. In a Convolutional Neural Network (or CNN as it will be described in the rest of this paper) we can expect a *sui generis* formulation of the model. Analogous to what has been revised, this network will also take various inputs, where most commonly the input is an image, then decomposed, so it has some computational meaning. The CNN relies on the fact that the input will consist mainly of images, which consequently will revise the whole structure of the model. Thus, unlike the classical artificial neural network, the CNN will arrange neurons typically in a three-dimensional manner, where they can also be one-dimensional. Typically, we will see inputs defined as  $width \times height \times depth$ . As a product of what we can assess, the layers composing this CNN must be constructed in a way to deal with the sensitivity of this kind of input, and generate an output according to the same conjecture. More often, CNNs, as depicted in Figure 2.4, are designed with the following layers: Input layer, Convolutional layer, Pooling layer, Fully-Connected layer, and Output layer.

Convolutional layers are the fundamental building block of CNNs. These layers perform feature

extraction, applying a kernel across the input, being an array of numerical values. Convolution operations take the input and apply the kernel, resulting in a feature map, which consists of the sum of the features on the input, after being multiplied by the kernel. The feature map is itself a matrix structure with dimensions defined by the number of times the kernel can query the input, applying the convolution operation. The pooling layers will take as input the feature map, extracting patches from this map, and reducing the dimensionality. The most common pooling operation is max pooling, where one takes as input a feature map, and retrieves the maximum value, to add to an output map. Other types of operations can also be utilized, such as average pooling. When reaching the fully-connected layer (or dense layer), the down-sampling of the pooling layer combined with the extraction performed by the convolutional layer will be measured by a probabilistic classification, much like a typical neural network. The reason behind CNNs being able to make a great amount of computation in parallel is due to the fact that each input can be computed at the same instant, basically when they are not interdependent.

The problem is that, in Convolutional Neural Networks, an image is derived from the notion that one pixel depends on its proximity, and the next pixel on its pressing neighbors, working upon them and applying filters on patches of an image, in order to gather relevant features. However, if instead of patches the model is given the whole image, the chances of improvement increase. This is one reason to support the use of transformers instead of CNNs.

## 2.5 Transformers

Transformers are taking the current Artificial Intelligence world at an incredible pace. Nowadays, we can expect these models to be present in numerous applications related to natural language procession and computer vision tasks. Some well-renowned models based on the Transformer architecture include BERT ([Devlin et al., 2018](#)) and GPT-2 ([Radford et al., 2019](#)). Transformers use self-attention modules, which according to [Vaswani et al. \(2017\)](#) boosts the speed of how fast the model can translate from one sequence to another.

As stated by [Vaswani et al. \(2017\)](#), the transformer is a model that resembles the most competitive models, since it relies heavily on an encoder-decoder design. In order to give the best explanation of what really is a transformer, firstly, it is advised to understand how the inputs are fed to this model, and what is an Attention mechanism. The input, given in any form, has to be modified, as transformers take only account of numeric structures. There is no notion of giving the full image, or a single frame, as neural models learn in accordance with numbers. Thus, the input has to be formulated in what is called an input embedding. This embedding can be visualized as a vector representation of what a word/patch really is computational. This vectorized input is then passed through a positional encoder in order to give each embedding an estimation weight according to its position on the input. The need for doing

such an encoding, relative to the position, is due to the fact of the lack of recurrence of the model. Thus, the authors (Vaswani et al., 2017) stated that this encoding should go according to the following two Equations 2.5 and 2.6, where  $n$  is a scalar defined by authors with the value of 10,000.

$$PE_{(pos,2i)} = \sin(pos/n^{2i/d_{model}}) \quad (2.5)$$

$$PE_{(pos,2i+1)} = \cos(pos/n^{2i/d_{model}}) \quad (2.6)$$

In short, the use of sine and cosine functions attends to the learning ease that they provide. Furthermore, it is noticeable that for an even position, the function used should be as defined by Equation 2.5, and for an odd position in the embedding, the function to use is defined in Equation 2.6 accordingly. These equations grant a sinusoidal representation of the feature embedding, as it can develop a single unique encoding for each position. Finally, the vector resulting from this operation should be summed to input embeddings, in order to give the model current information on the positioning of the vectors. This encoder is mainly composed of two operations: multi-head attention, and a feed-forward network. Regarding the attention mechanism, it enhances the performance of the model by searching a set of positions in the encoder states to find the most suitable/relevant data, when some generation has taken place. To understand this more deeply, let us look in more detail at the figure based on the work done by Bahdanau et al. (2014), as seen in Figure 2.5. The authors Vaswani et al. (2017) propose a Multi-head architecture, whereas there can also be Scaled dot-product attention, as can be seen in Figure 2.6.

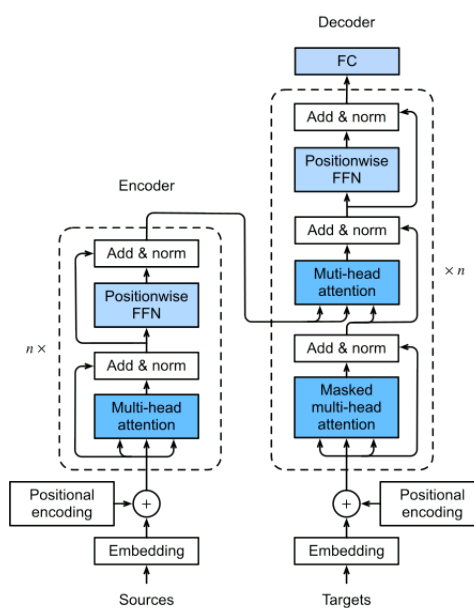


Figure 2.5: Transformer model following the specifications of Vaswani et al. (2017).

Regarding the Scaled dot-product attention method, there are three main components needed: the Query vector, the Key vector, and the Value vector. These vectors are then multiplied for matrices created during training. Having that  $d_k$  represents the dimension of the Keys, every word will have one of each, and the matrix outputs are calculated as follows.

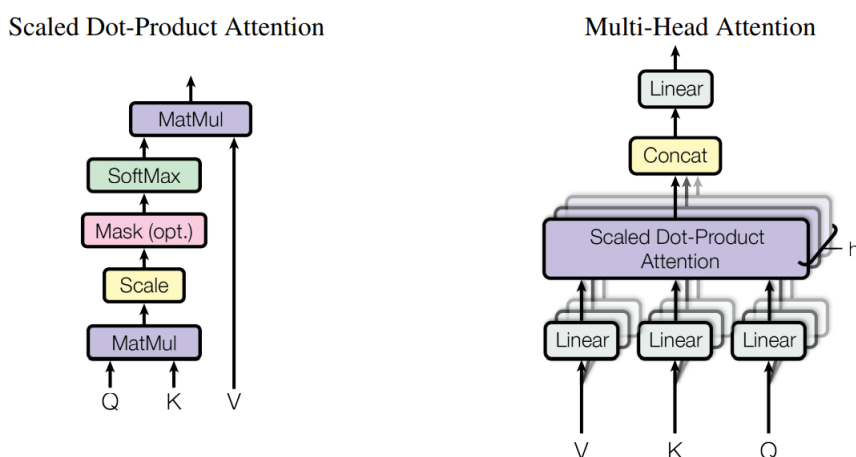
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.7)$$

In order to counter-measure the small gradients of softmax, due to the increased complexity of the scaling factor resulting in the dot product of  $Q$  with the transposed matrix  $K$  (for a large enough  $d_k$ ), the authors suggest that the result should be divided by the scaling factor  $d_k$ . Although results were fair, concerning Vaswani et al. (2017), it was devised a new way of performing the attention method, by linearly projecting the query, keys, and values vectors  $h$  times. Then, in parallel, perform the attention Function 2.9 to each linear project vector. This method is called Multi-head attention.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2.8)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2.9)$$

In Function 2.9, the projection matrix for each of the vectors is stated as  $W_i^Q$ ,  $W_i^K$  and  $W_i^V$ . This method enhances the attention search on less deterministic spaces. Consequently, the single attention employing the average diminishes this kind of result. From this attention submodule to the fully connected feed-forward there are channels for passing the outputs as well as the normalization vectors. For both the encoder and decoder, there is a fully connected feed-forward network with a ReLU activation



**Figure 2.6:** Single attention and Multi-head attention, according to Vaswani et al. (2017).



function.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2.10)$$

The outputs from this final stage of the encoder will pass to the decoder. The decoder, as is defined in Figure 2.5 is formulated in order to generate the text sequences. These sub-layers behave very similarly to the encoder layers, but each multi-head attention layer performs a different task. It also has position-wise feed-forward layers, residual connections, and normalization layers after each sub-layer. Also, the decoder is topped off with a linear layer that functions as a classifier followed by a softmax function, to determine the words' probabilities. It begins the decoding process with a start token, and it takes in a list of previous outputs, as well as encoder outputs, which contain the attention information from the inputs. After generating a token, the decoder stops the decoding process

To obtain positional embeddings, the input goes through an embedding layer and a positional encoding layer. In the first multi-head attention layer, the positional embeddings are used to compute the attention scores for the decoder's input. This multi-headed attention layer works uniquely. There must be a prevention mechanism for the decoder not to condition future tokens since it is auto-regressive and creates the sequence word by word.

Thus, a look-forward mask is used to restrict the decoder from looking at future tokens. The mask is applied before the softmax is calculated and after the scores are scaled. This is a matrix the same size as the attention scores that is filled with 0s and negative infinities. When the mask is applied to the scaled attention scores, the result is a matrix of those, with the top right triangle filled with negative infinities. The mask is used because once the softmax of the masked scores is reached, the negative infinities are wiped out, leaving zero attention scores for subsequent tokens. This instructs the model to ignore those words. The only change in how the attention scores are calculated in the first multi-headed attention layer is the masking. This layer continues to have numerous heads to which the mask is applied before being concatenated and passed via a linear layer for further processing. The initial multi-headed attention produces a masked output vector with information on how the model should attend to the decoder's input

The last position-wise feed-forward layer's output is routed through a final linear layer that serves as a classifier. The classifier is as large as the number of classes. For example, if you have 10,000 classes for 10,000 words, the classifier's output will be 10,000. The classifier's output is then input into a softmax layer, which generates probability ratings ranging from 0 to 1. We choose the index with the greatest likelihood score and multiply it by the anticipated word. The decoder then adds the output to the list of decoder inputs and continues decoding until a token is expected. In our situation, the final class assigned to the end token has the highest probability of prediction.



# 3

## Related Work

### Contents

---

3.1 Report Generation Based on Transformers . . . . .	21
3.2 General Image Captioning . . . . .	27
3.3 Retrieval Mechanism in Image Captioning Tasks . . . . .	31
3.4 Overview . . . . .	34

---



In this related work, many of the works presented can also be seen in the survey by [Litjens et al. \(2017\)](#), where the authors themselves propose some methods that make out the state-of-the-art panel of the method to this day. For this section, the proposal is similar, it is done as an overview will, of the work and investigation that has already been done on the matter presented. It's proposed to evaluate these works separately and then provide an insight into what is pertinent to the objectives of this work. Thus, this section is separated into each of the papers that emphasize the magnified performance of CNN and Transformers (combined), and then, works that use only Transformer based methods. Also, there will be some related work concerning methods and datasets specifically enhanced to achieve better performance when training such models. Moreover, some works on the impact of using a retrieval mechanism will be presented, as without them this dissertation would not have reached such results. This chapter is organized such that each section has details about each and every relevant work.

## 3.1 Report Generation Based on Transformers

The focus of the study will be on architectures that are mainly formed by Transformers. This is more detailed in the current work, as the purposed architecture will be of the same topology. Also, it is relevant to contrast the main differences in performance and accuracy when dealing with CNN+Transformers and only a Transformer-based model.

### 3.1.1 Generating Radiology Reports Via Memory-Driven Transformer

One proposal for solving the problem using Transformers was indicated by [Vaswani et al. \(2017\)](#). For trying to improve already developed architectures, the authors use a memory-driven transformer shown in Figure 3.1, where the use of the Relational Memory (RM) is set to save previous generations and a memory-driven conditional layer normalization (MCLN) ([Chen et al., 2020](#)), to incorporate this memory to the transformer. They impose the sequence-to-sequence paradigm, where, using a visual extractor, they feed the patches from the source image and tokens from the generated report, according to the X-Ray. Thus, assessing that, there will be a need for a visual extractor. The current model presented by the authors follows a structure very much like a common transformer, where there is an encoder-decoder architecture plus the visual extractor, the memory-driven conditional layer normalization, and the RM.

In the visual extractor, the idea is to sort patches of the radiology image, from a pre-trained CNN. Finally, the encoded sequence is used as the main source for the encoder of the transformer. The encoder is used as defined in a typical transformer-based model, where the outputs must be the hidden states generated from the input. In spite of the memory-based architecture, the decoder had to be modified in order to adapt the Transformer to bring in the MCLN.

Regarding relational memory (one of the key aspects of this paper), this mechanism is used so that

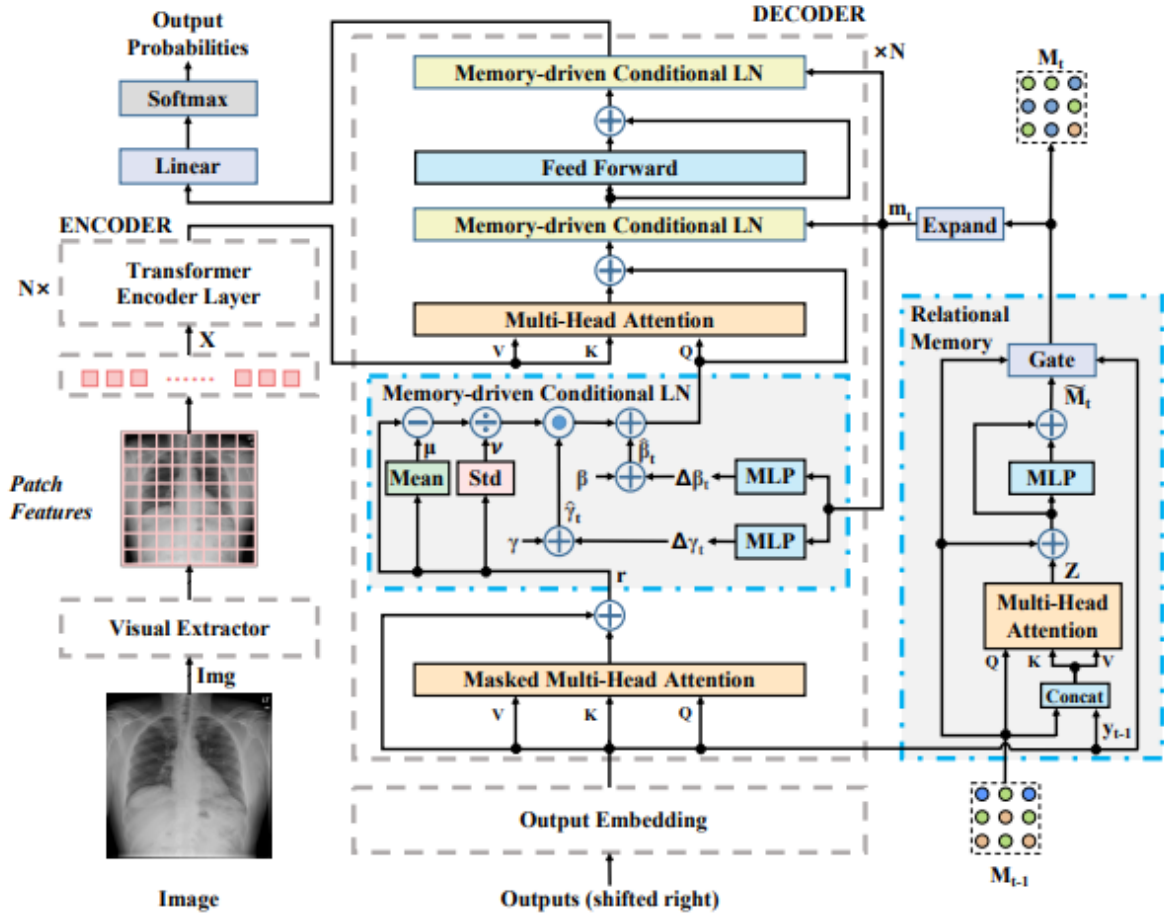


Figure 3.1: Architecture of a Memory-driven Transformer RM+MLCN (Chen et al., 2020).

when presented with one radiology image where the features are relevant, they include any other image that is somehow similar, in order to cross-reference the reports. Thus, the learning method can be improved, taking advantage of patterns on both reports from similar radiology images. The practical way that Chen et al. (2020) do this, is by creating a matrix where each row must define the pattern's data. The main idea is that the matrix can be updated in each generation, with the results from the previous ones. Regarding the proposal made to incorporate the MLCN module in accordance with a previously defined model of the Transformer, the authors justify this incorporation with the fact that text generation is a procedure that requires dynamic management of the decoding output per generation. The MLCN, therefore, will act as a linking bridge between the RM and the rest of the decoder. To test the impact of these choices, the authors made tests using both MIMIC-CXR (Johnson et al., 2019) and IU X-RAY (Demner-Fushman et al., 2016), and test the Transformer with the use of the RM and both RM+MLCN. Using the IU X-RAY dataset, the results show that the improvement over the baseline model goes beyond 8.9% with only a Relational Memory module. With both Relational Memory and Memory-

driven conditional layer normalization, the improvement overcomes the latter with an average of 17.6%, which is an improvement of 8.7% over the model with only a relational memory. Under the MIMIC-CXR dataset, the results confirm that, again, using the RM+MLCN model will improve the precision with an astounding percentage (a difference of 8.4%).

### 3.1.2 Progressive Transformer-Based Generation of Radiology Reports

Interestingly, [Nooralahzadeh et al. \(2021\)](#) has given another perspective on the task of generating accurate X-Ray reports, given front-side and lateral imagery of a thorax. They suggest splitting the process into two parts: one is an intermediate result of an encoder-decoder for a Visual Model, and the final steps will consist of an encoder-decoder for a Language Model. Particularly, the first step will generate high-level concepts, to create more detailed text sequences given a context.

The proposal is defined as image-to-text-to-text, as seen in Figure 3.2, following the previous definitions of the steps taken by the given approach. The idea here is to propose an intermediate step for generating the context needed according only to the image. More often than not, the generating pass immediately for the whole report. Here, the creators have a Visual Language Model that takes as input two vectors, each one according to the image features extracted by the Visual Backbone (based on a CNN). These vectors go through the ViLM, being a Meshed-Memory Transformer, which implies that the encoder will act in order to gather prior information and store it to later pass it to the decoder. The decoder, being a meshed decoder, the decoding will provide a cross-attention between all encoding layers and the decoder layers. Finally, the Language Model (employing a pre-trained transformer BART), is used for being an autoregressive decoder capable of generating sequences with paraphrasing and summarization.

To train the architecture, the authors use a state-of-the-art method for extracting entities and rule base negation detection with MIRQL, proposed by [Irvin et al. \(2019\)](#). Also, the datasets used were the IU X-RAY ([Demner-Fushman et al., 2016](#)) and MIMIC-CXR ([Johnson et al., 2019](#)), with the BLEU ([Papineni et al., 2002](#)), METEOR ([Lin and Och, 2004](#)) and ROUGE-L ([Banerjee and Lavie, 2005](#)) metrics. The

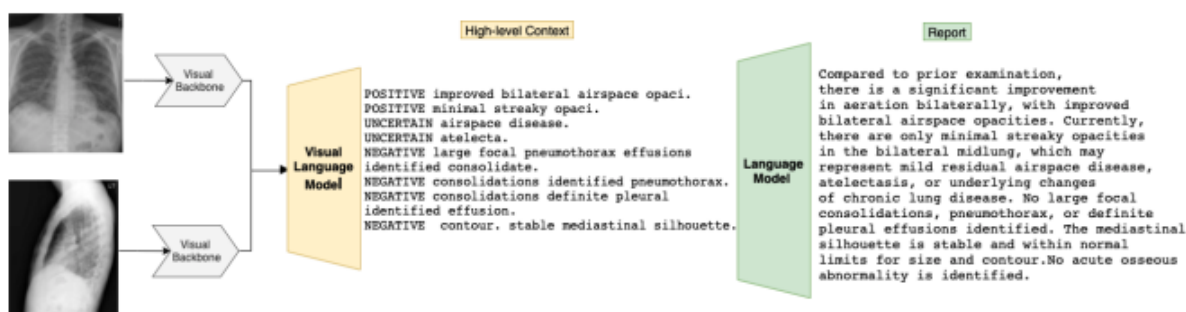


Figure 3.2: Framework of the  $M^2$  Tr. Progressive ([Nooralahzadeh et al., 2021](#)).

results provided by the experiments show great promise in the provided method against the baselines. Furthermore, although results are so promising, work on coherence is still needed.

### 3.1.3 Learning to Generate Clinically Coherent Chest X-ray Reports

The work of [Lovelace and Mortazavi \(2020\)](#) proposes another automated radiology report generation model, where the use of the Transformer architecture is crucial to correct the incorrect notions that the models should only be preserved on the NLP metric basis. The authors of this paper propose to also evaluate the model on clinical accuracy metrics such as precision, recall, and the F1 score. It is typical, even in past works, to test only according to the fluency and human text correlation. However, when providing models to solve clinical problems, we also add another problem to the equation, being that the clinical accuracy, and is not solved by just evaluating the NLP performance.

The model presented in this work is very similar to the last ones presented in this same related work. Also trained and tested in the MIMIC-CXR dataset ([Johnson et al., 2019](#)), this model has been constructed in such a way that the generation has two crucial steps. Firstly, the authors create a model that will represent reports by simply training the model on standard language generation. Secondly, the model will extract the most clinically relevant features from the observation made by the encoder, and compare them with ground-truth reports, where this additional step may increase the clinical coherence.

This Transformer designed as Clinical Transformer, uses a DenseNet-121<sup>1</sup> model, for extracting the visual features, whereas the decoder is based on the neural machine translation model (NMT) by [Chen et al. \(2020\)](#).

The results, in what concerns the language correlation with the actual reports, show that the model has surpassed the compared models in all NLP metrics. However, this does not infer anything from the clinical accuracy standpoint. As for this, the authors have tested the model on macro- and micro-averaged precision, recall, and F1 score, showing that compared to the late CNN model, all of these metrics have been surpassed by the Clinical Transformer.

### 3.1.4 Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation

The following model, denominated as Posterior-and-Prior Knowledge Exploring-and-Distilling (PPKED), was proposed by [Liu et al. \(2021b\)](#), who try to mitigate the data bias that is present in data-driven models. Given that, PPKED is created to mirror the process of diagnosis that radiologists use to write their clinical reports. In the first instance, the radiologist searches for abnormalities in the x-ray, and then, given those abnormalities, they indicate possible diseases, finally relying heavily on prior knowledge.

---

<sup>1</sup><https://pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html>



This transformer-based model is composed of three modules: Posterior Knowledge Explorer (PoKE), Prior Knowledge Explorer (PrKE), and Multi-domain Knowledge Distiller (MKD). In short, the PoKE model will be responsible for the posterior knowledge withdrawal from the input image (i.e. anomalous areas of the x-ray). Then, the PrKE will assert knowledge from past reports, retrieving as much information as it possibly can, in order to increase the report corpus and better represent the medical components. Finally, the MKD module is where all this prior and posterior knowledge will be input, using a text decoder to generate a clinical report as close to the prior knowledge as possible.

To evaluate this methodology, the authors [Liu et al. \(2021b\)](#) have tested the model in both the MIMIC-CXR ([Johnson et al., 2019](#)) and IU-Xray ([Harzig et al., 2019](#)) datasets. One of the things that PPKED has left unchecked is the lack of attention regarding clinical efficiency, having no regard for precision or recall, contrarily to the Clinical Transformer ([Lovelace and Mortazavi, 2020](#)) and MDT+WCL ([Yan et al., 2021](#)). However, concerning the NLP basis and metrics such as BLEU ([Papineni et al., 2002](#)) and ROUGE-L ([Lin and Och, 2004](#)), the PPKED model has overcome performance in comparison to all other models (i.e. CNN-RNN ([Vinyals et al., 2014](#))). Furthermore, the highest achieving ROUGE-L score in their study was over 27.7, whereas PPKED managed to achieve 28.4 in the same metric.

Although results show that PPKED is achieving promising performance when compared to well-established models, the lack of study in what concerns the clinical efficiency of the model can leave these scores to mean less than what they present.

### **3.1.5 Automated Generation of Accurate & Fluent Medical X-ray Reports**

Focusing again on the issue of creating human readable reports over clinically accurate ones, the authors of this work ([Nguyen et al., 2021a](#)) have focused their attention on the creation of a model that, by taking both radiology images and the clinical history of the patients as input to a classification module, where the diagnosis will be put into topics, where each is a disease. Having this list of diseases is then passed as, what the authors call an enriched disease embedding, where the model can generate the medical reports.

This model is defined as a three-phase framework, where they present a classification module, a generation module, and finally an interpretation module. The first module (classification) will have numerous x-rays as inputs, where they have to extract the visual features, relying on an image encoder. Furthermore, the text encoder will extract the most pertinent clinical features from the report document. Like our work, they concatenate both vision and text features into a single embedding, using an "add layerNorm". Then, the generation module takes this embedding and generates the report token by token. This proposed report is finally passed into an interpretation module, where they assess the clinical accuracy achieved. This is used to keep the model as accurate as possible, having the list of diseases as a checklist where the model verifies the proposed report over this enriched disease embedding.

### 3.1.6 Weakly Supervised Contrastive Learning for Chest X-ray Report Generation

This proposal by [Yan et al. \(2021\)](#) is based on a weakly supervised contrastive learning framework. This work demonstrates an approach close to our own, where the authors have shown that the final report benefits from contrasting the query report with similar ones, although they might be incorrect. The method they define as MDT+WCL uses ChexBERT ([Irvin et al., 2019](#)) to label reports and use them as a weakly supervised signal after the contrastive and cross-entropy functions. For this, they use a well-known memory-driven transformer ([Miura et al., 2020](#)) as the baseline model. To extract the visual features they use the pre-trained convolutional neural network ResNet (cite).

In their experiment, they not only test on NLP metrics but on clinical accuracy as well. Comparing their approach with similar models, both on MIMIC-CXR and MIMIC-ABN, it is clear that MDT+WCL can outperform past models, providing a boost in clinical efficiency with their contrastive loss. This model has also the capacity of developing reports with an accurate diagnosis of anomalous findings, which is a setback for most models, as they will train on large datasets, where the abnormal findings will be scarce, and will tendentially generate the most common clinical features.

The results presented show that their model outperforms all other models (e.g. ([Liu et al., 2021b](#))). Furthermore, the performance is not only visible in the NLP metrics, as they grant a better performance concerning the clinical accuracy of their reports. This sets a new boundary to the Transformer-based model, in the report generation task. However, their model does not consider any specific details on the location of a disease, whereas it can correctly diagnose a pleural effusion, but it is not clear where it is located.

### 3.1.7 Aligntransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation

Another proposal by [You et al. \(2021\)](#) is introduced to mitigate problems in medical report generation. Problems such as data bias and long sequences. To tackle these issues the authors have introduced a model called AlignTransformer, where this model consists of an Align Hierarchical Attention (AHA) and Multi-Grained Transformer (MGT) modules. In what concerns the AHA module, will predict any clinical issues that may appear in the radiology image, finally learning these multi-grained visual features sets up them into a hierarchical alignment. The second module (MGT) will be responsible for using these features to generate the medical report. Furthermore, this MGT module exploits a similar technique as the Memory-meshed transformer ([Chen et al., 2020](#)), by using the memory-driven conditional layer normalization (MLCN). Similarly to the past models and our own, this AlignTransformer is trained and tested on both MIMIC-CXR and the IU-XRAY.

According to the ablation study, where the authors assess the performance that this model might introduce in the state-of-the-art, it showed that the use of an AHA module will increase the performance on all NLP metrics, achieving a gain of 18.8% in BLEU-4, over the determined baseline. As for the MGT, it was visible that it can provide an additional capacity to adaptively search the multi-grained visual features.

## 3.2 General Image Captioning

In this section, we will a broad spectrum of models that were introduced as image caption architectures. We will further develop further on the impact these methods provide on such tasks, and how they work compared to each other. It is important to know what is expected of such models, regarding a task that is less demanding than clinical report generation.

### 3.2.1 How Much Can CLIP Benefit Vision-and-Language Tasks?

As, image captioning is a task of great importance to the current project, concerning the prediction of text that conveys the meaning of given image input. For this purpose, the work done by [Shen et al. \(2021\)](#) goes far beyond the usual visual encoder. Often, visual encoders are trained upon short-scaled datasets. However, the proposal is to, indeed, validate the use of a pre-trained visual encoder to be used with large-scale image-caption pairs. Moreover, the authors combine CLIP (Contrastive Language-Image Pre-training) for fine-tuning specific tasks, and also, the combination of this encoder with V&L pre-training. The key point of using CLIP is that for any visual encoder if CLIP performs according to expectations, CLIP can be set as the visual encoder.

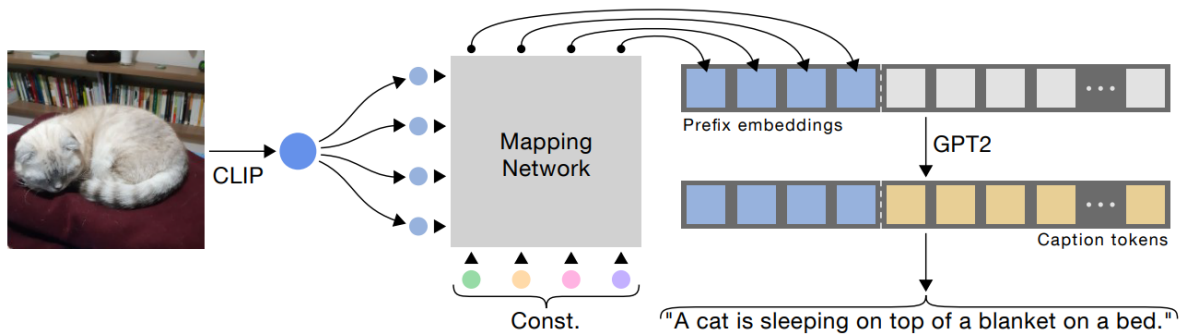
Two scenarios are presented, CLIP-ViL and CLIP-ViLp, where the first recurs to pre-training, contrarily to the latter. Upon these proposals, three main tasks of V&L are tested with the use (and absence) of CLIP. These are Visual Question Answering (VQA), Image Caption, and Visual-and-Language Navigation. Although two of the tasks are predominant in V&L tasks, the focus will be increasingly directed to Image Caption impact due to CLIP incorporation. The way CLIP was devised was such that, following a shallow-interaction design, to a given visual and text encoder performs the task on the image and text independently, resulting on a dot-product between both encoders, generating the output results, stated by the authors as *Alignment Result*. Also, the pre-training of CLIP consists of more than 400M image-text pairs.

Regarding the image caption task, CLIP-ViL was adapted onto a Transformer (very much like this project). The performance of this adaptation, with use of COCO dataset ([Lin et al., 2014](#)), is represented by the METEOR ([Lin and Och, 2004](#)), CIDEr ([Vedantam et al., 2015](#)), BLEU ([Papineni et al., 2002](#)), and SPICE ([Anderson et al., 2016](#)) metrics. Results from this experiment show that CLIP-Res50x4

is capable of outperforming SOTA models. Moreover, with the increasing complexity of model size, it also performs better than ResNet50 and ResNet101. However, the model aforementioned does not perform any pre-training. This might be costly to the efficiency of the visual encoder, as pre-training has been shown to leverage the performance of V&L tasks. Having said that, the CLIP-ViLp model is set with word embeddings, resulting in the tokenizations of text, using the sum of the position and the segment embedding. On the image side, the embeddings are formulated as a series of vectors acting like a feature map, where they are later concatenated to the word embeddings. The results, after this pre-training, show considerable performance gains on all tasks, but GQA. Proposing to use CLIP as a replacement for a visual encoder on V&L tasks can show great promise in the overall model performance. Although these results are positive, some work must also be inquired to detect where to draw the line between CLIP and other SOTA proposals. One key aspect that could be retrieved from this work ([Shen et al., 2021](#)), is that localization can be an issue for this encoder, shown as a result of Grad-CAM experiments done on the CLIP variants. Another point worth to mentions is the possibility of unfreezing the Visual Backbone, as is shown to be capable of enhancing the performance of the CLIP-Res50 and CLIP-Res50x4.

### 3.2.2 ClipCap: CLIP Prefix for Image Captioning

Still following the line of work done on CLIP, the work presented in [Mokady et al. \(2021\)](#) suggests the incorporation of CLIP in order to correctly caption a given image, as seen in Figure 3.3. In accordance with an already pre-trained language model (GPT-2 in this case, due to the richness of the texts provided), the proposal employs this model with the CLIP, formulating, therefore, ClipClap. Moreover, this conjunction of models will try to tackle two main problems with image captions: one concerns the best way to describe the image since there are considerable amounts of syntactic and vocabulary combinations to define a text. Another issue, typical of such works, is the semantic involvement of the model. For example, for any given image, the caption should be such that no information is missing, and no information is falsely describing the scenario. So, ClipClap will consist of the use of CLIP encoder ([Shen et al., 2021](#)), where there is the possibility of leveraging the representation of the image and textual references, correlating both. On a more detailed instance, the work done by [Mokady et al. \(2021\)](#), will generate prefixes for each of the textual captions over a mapping among the CLIP embeddings. These prefixes are embeddings of fixed size, concatenated to caption embeddings. It is important to mention that the results provided by ClipClap are, in the majority, the product of a Transformer-based architecture, creating more meaningful captions with the least demanding parameter model. One of the imposed difficulties in using CLIP with GPT-2, is the need of understanding that, representations for both models differ in the representation of text. For this, the authors employ the fine-tuning of the language model during the training of the mapping network. This, they state, can employ better flexibility and generate



**Figure 3.3:** Overview of ClipClap transformer-based architecture.

more meaningful outputs. On the other hand, the translation of CLIP embedding to the language model space is maneuvered with the help of the multi-attention mechanism provided by the transformer. As the model is fed with the visual encoding from the CLIP and a learned constant, simultaneously, more information is retrieved from the CLIP embedding and is adjusted by the language model to more recent data.

In practical terms, the proposal will consist of the retrieval of a prefix from a visual context (provided by CLIP), generating the caption in accordance. Then, each token will be predicted one by one with help of GPT-2. Having the probability associated with each token, using beam search or greedy search, the more suitable token is selected (as seen in past approaches). Even so results haven't surpassed SOTA models, there must be no mistake in the capabilities presented by this proposal. The focus was to present an easy-to-use architecture that should not require heavy-duty training cycles of too much training focus.

### 3.2.3 CPTR

In another proposal to tackle this challenge of generating accurate reports with reference to a radiology image, the authors [Liu et al. \(2021c\)](#) have another vision of how to surpass some of the difficulties imposed. The introduction to the Caption Transformer (CPTR), shown in Figure 3.4, comes as a comparison to the CNN+Transformer systems, where the results have shown to be very promising. Also, to test the viability of this method, the authors work with MSCOCO dataset ([Lin et al., 2014](#)), which is a widely used dataset with common object images. Normally, the Transformer architecture is very much like the one that was referenced in Section 2.5. Instead of using a well-known method to gather the most predominant features from the input image, instead, they treat each image as a sequence-to-sequence task. Therefore, patches from the images are made, in order to create an input embedding. The encoder steps are followed exactly as depicted in Figure 2.5 where the inputs are a vectorized version of the image, where this image is really a patch from the original. Also, instead of using the ReLU activation

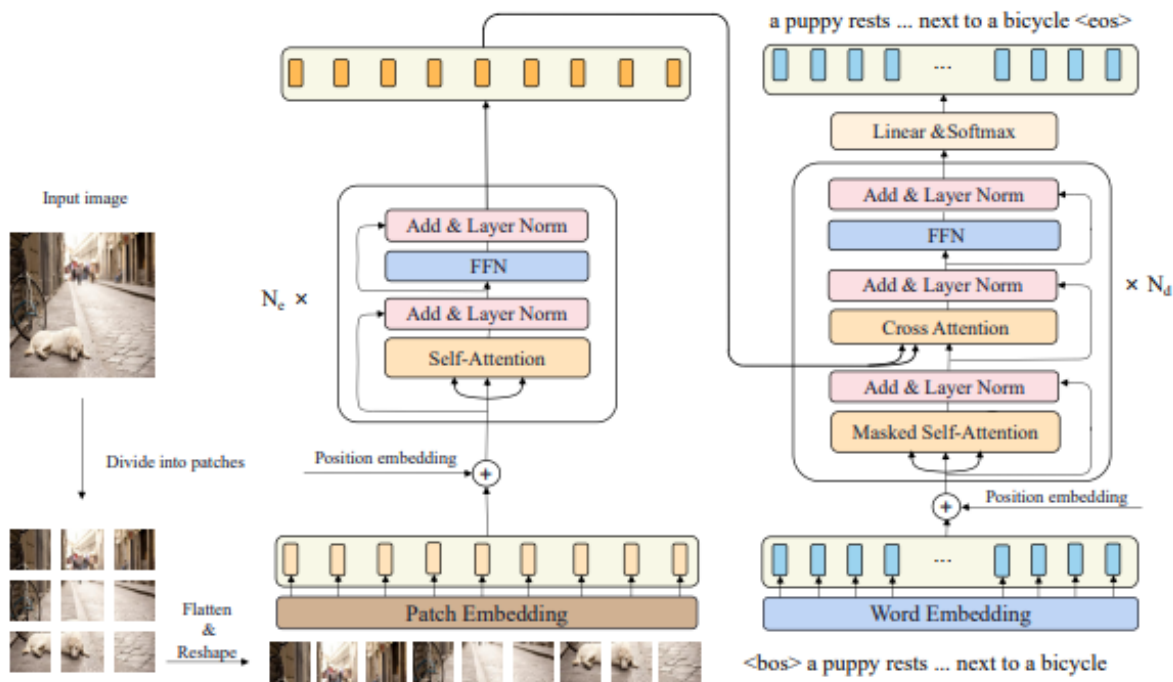


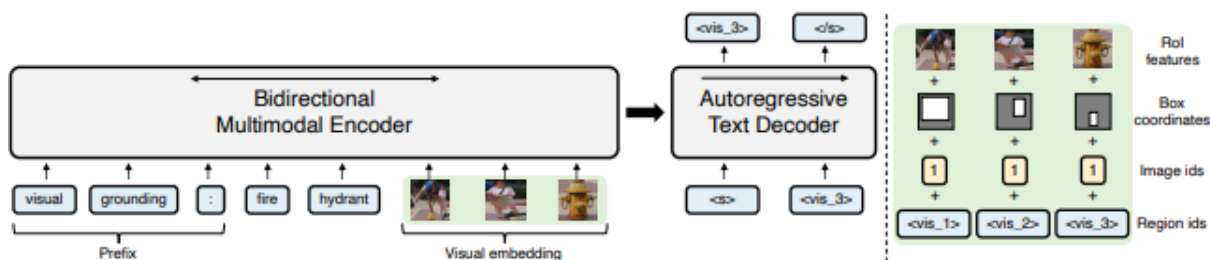
Figure 3.4: Architecture of CPTR (Liu et al., 2021c).

function, the creators of CPTR use GeLU (Gaussian Error Linear Unit) with a dropout factor, instead.

To test the CPTR against the CNN+Transformer and CNN+RNN paradigms, they used BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Lin and Och, 2004) and ROUGE (Banerjee and Lavie, 2005). With a Transformer that contains 12 layers on the Encoder module and 4 layers on the Decoder module, with 12 heads of attention on both the encoder and decoder. According to the results, the CPTR outperforms both paradigms, achieving a score of 129.4 on CIDEr, over 128.3 (CNN+RNN) and 127.8 (CNN+Transformer). These results are very interesting, as they elevate the relevance of migrating to Transformer methods to increase the reliability of image captioning, and replacing the most common methods, such as CNNs with Transformers.

### 3.2.4 VL-T5 and VL-BART

Following the previous paper on image captioning (Liu et al., 2021c), as it might be suggestive, there are numerous other approaches that promise the same kind of results, without the need of using convolutions to address the challenge. This is the example of Cho et al. (2021). This work introduces VL-T5 and VL-BART shown in Figure 3.5, established upon previously pre-trained Transformer models. This paper proposes a hypothesis to override the necessity of the excessive pre-training that a new task demands. In order to validate this, the authors Cho et al. (2021) make vast tests to 7 tasks. Firstly, on the Visual



**Figure 3.5:** Architecture of VL-T5 and VL-BART (Cho et al., 2021) for visual grounding task.

QA task, the authors use two datasets, VQA and GQA. This specific task calls for a model able of answering questions given an image context. Then, NL Visual Reasoning is a binary task, responsible for stating if any given statement is generated if it correctly describes any of the images presented. On the Referring Expression Comprehension task, the model should try to identify the patch of the image that better represents any queried object. VCR (or Visual Common-sense Reasoning) requires that for a set of 4 rationales and 4 questions, the model can correctly select the o correct answer and rationale. The Multimodal Machine Translation tries to figure out a new approach to translate a caption of any sentence to another language, given two modalities, the image, and the caption. In this specific study, the authors tried the translation from English to German. Finally, and a more centered task on what regards this project, the Image Captioning task used upon the MSCOCO (Lin et al., 2014) dataset should generate captions for a given image, as we have seen in the previous papers. Regarding the model presented in the work by Cho et al. (2021), it is used in a visual embedding that actually defines regions of interest on the image, acting as a feature extraction technique for identifying points of interest. Following the encoder, the Auto-regressive Text Decoder will generate the sequence of text according to the text input. In this paper, the visual embedding is concatenated with the text embedding, acting as a text-to-text generation, unifying the image with the text.

### 3.3 Retrieval Mechanism in Image Captioning Tasks

In this final section of our related work, we present some models that use a retrieval-based mechanism, in order to enhance the performance of image captioning architectures. This idea comes simply by retrieving information stored, an already known, to later use as a guide for improving the generation. Furthermore, we will give special attention to these selected works, since they provide solid ground for this dissertation. Proposals like submission (2022), is very aligned with our own proposal, although the tasks at hand are similar, but not so similar after all.

### 3.3.1 Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model

In this proposal, the authors [Endo et al. \(2021\)](#) introduce a model denominated as CXR-RePaiR, being a retrieval-based radiology report approach to enable the betterment of generation performance. This one is especially relevant as it brings up many of the components that this project will propose. One of these components is CLIP ([Shen et al., 2021](#)), which has already been determined as a versatile instrument for image and text encoding/decoding. Furthermore, in order to take performance to a more optimized instance, the authors also propose the compression of the corpus, making it lighter for the creation of new generations of captions. As is clearly shown by the architecture of the system, the proposed model CXR-RePaiR will take a corpus of reports and pass them through the text encoder of CLIP, creating the mapping for each one of these. Also, an x-ray will be fed to the image encoder, again creating the mapping for the image in question. Then, by CLIP execution, a mapping of both image and text will be formulated. This conjunction made, will generate a possible caption for the combination made by CLIP. This generation happens with the selection of the report that maximizes the similarity between the embedding previously created. Then, the predicted report is inputted onto CheXbert, where F1 scores are set between the prediction and ground truth.

For enabling better higher dot-product similarity values between image pairs, pre-training phases are set in motion using natural image pairs, and then report-image pairs. Resulting also, in low dot product between pairs of separate instances. To compare the performance of this model against some SOTA models, they select the M2 Trans and the R2Gen, which was already discussed in this same project. As textual components, the model uses both sentences from the reports, and also the reports. This aims for improving the generation of unforeseen reports without the need of combining full reports in the prediction process. Furthermore, to create a more sustained base for this to happen, the model will use MIMIC-CXR as the internal dataset and CheXpert as the external dataset. However, this leaves a question as to how are the sentences chosen for the generation. To circumvent any biased selection, the number of sentences that can be retrieved has to be directly influenced by the contents of the predicted sentences, using the function  $\text{argmax}_{s \in S(R)} f(x, y)$ , where  $S(R)$  are the sentences returned. In the first set of experiments, the authors figured that CXR-RePaiR-Select scored higher than any other on the F1 score ( $0.274 \pm 0.003$ ). Also, in a more contextual experiment, the authors compare reports for each of the SOTA models against the improved CXR-RePaiR (CXR-RePaiR-3). What can be visualized from the results is that the proposed model has more positive predictions while improving the readability of the report itself. Furthermore, the report presented with CXR-RePaiR-3 presents almost exclusively clinically useful information, disregarding any arbitrary components. Although these results show that retrieval can be a major improvement in image captioning (in focus on the clinical theme), there are some considerations taking place in this proposal. The fact that only two SOTA models are being compared



against the model CXR-RePaiR leaves some space for questioning the reliability of the results against other proposals since the collection of results is very limited. On another point, the comparison against models that also propose a retrieval base mechanism can be enough to show that, even among the best models, CXR-RePaiR has shown great promise. However, validating the model against models with different strategies for improving the captioning task would be ideal. Another consideration on a more detailed aspect of CXR-RePaiR is the lack of training for rare diseases. As they propose a training process that selects sentences regarding the maximum score according to components of select reports, this leaves the more anomalous findings to be somehow disregarded.

### 3.3.2 Retrieval-Augmented Image Captioning

The work proposed by [submission \(2022\)](#) also tries to leverage a Transformer-based model to achieve better results with the addition of retrieval mechanisms. In this specific work, the Transformer is a multi-modal VL BERT encoder with a GPT-2 decoder. EXTRA, as they denote their model, stands for Encoder with Cross-modal representations Through Retrieval Augmentation, and works with a specific kind of similarity metric designed by Facebook. This similarity mechanism is the base of the retrieval mechanism, and it is called Facebook AI Similarity Search (FAISS) [Johnson et al. \(2017\)](#).

The model is trained and tested upon the MS COCO dataset, where the model when receiving a new image in the encoder, firstly searches on a datastore the k nearest representations to that query image. Given that representations, the retrieval mechanism will sort the k captions from each of those images with higher similarity to the query. Given that, such captions are then concatenated with the visual representation of the image, very similar to our proposal.

The results presented in this work by [submission \(2022\)](#) lead us to believe that the use of a pretrained VL model, equipped with a retrieval mechanism will make the whole task benefit from it. When compared with models such as CPTR ([Liu et al., 2021c](#)), EXTRA has shown competitive results, improving the CIDEr results by over 2 points.

### 3.3.3 Retrieval-Augmented Transformer for Image Captioning

Another given proposal ([Sarto et al., 2022](#)) to increase the reliability of transformer models, relies again on the use of retrieval mechanisms. This work focuses on image captioning tasks while using a model capable of combining both the visual features and using them to achieve the most similar image from a memory bank. By acting upon this memory bank, working as a K-nearest-neighbors function, they retrieve the captions from the top-k similar images.

The results of this work have shown that Transformer-based models can gain with the use of external memory. By assessing that most of the captions retrieved match the query, they can clearly state that

Model	Author(s)	Task
CPTR	<a href="#">Liu et al. (2021c)</a>	General Image Captioning
VL-T5	<a href="#">Cho et al. (2021)</a>	General Image Captioning
VL-BART	<a href="#">Cho et al. (2021)</a>	General Image Captioning
CLIP	<a href="#">Radford et al. (2021a)</a>	General Image Captioning
ClipCap	<a href="#">Mokady et al. (2021)</a>	General Image Captioning
Clinical Transformer	<a href="#">Lovell and Mortazavi (2020)</a>	Automatic Report Generation
CXR-RePaiR	<a href="#">Endo et al. (2021)</a>	Automatic Report Generation
Transformer w/ RM	<a href="#">Vaswani et al. (2017)</a>	Automatic Report Generation
Transformer w/ RM + MLCN	<a href="#">Vaswani et al. (2017)</a>	Automatic Report Generation
$M^2$ TR.	<a href="#">Nooralahzadeh et al. (2021)</a>	Automatic Report Generation
$M^2$ TR. Progressive	<a href="#">Nooralahzadeh et al. (2021)</a>	Automatic Report Generation
PPKED	<a href="#">Liu et al. (2021b)</a>	Automatic Report Generation
Aligned Transformer	<a href="#">You et al. (2021)</a>	Automatic Report Generation
Nguyen et al.	<a href="#">Nguyen et al. (2021a)</a>	Automatic Report Generation
MDT + WCL	<a href="#">Yan et al. (2021)</a>	Automatic Report Generation

**Table 3.1:** Summary of the models presented in the related work section.

the level of adequacy using such retrieved information will not deform the performance already reached by such a model. Furthermore, the content of the retrieved sentences, depicting very close content to the caption of the query image, grants a more detailed caption, as the model without such a mechanism, fails to provide enough detail.

### 3.3.4 Memory-Augmented Image Captioning

The basis of deep learning methods for image captioning tasks has been recorded as successful, within the parameters of practical achievements. However, expanding the capabilities of such models on providing more accurate and fluent captions is not as trivial as it might look. To tackle this issue, using approaches with attention to retrieval-based memory mechanisms can be seen as a way to provide already-known information, to better guide the generation.

Such works like [Fei \(2021\)](#), leverage image captioning tasks by introducing a retrieval-based memory mechanism, where they interpolate the next word distribution with top-k matches. This will allow for a recall of already known information within a bank of data, similarly to our proposal. The experiments done upon this proposal by [Fei \(2021\)](#), have shown that upon the MS COCO benchmark, has proven that this memory-augmented mechanism can improve the caption quality, as well as, keep the capacity of leveraging other models that use fixed-size context representations.

## 3.4 Overview

To summarize the most crucial models that offer ground knowledge for this dissertation, we organize each one of these in Table 3.1, where we can state what problem they propose to solve.

In this section, we explored in the first instance, methodologies that try to improve the general image captioning task. For this, we have models like CPTR (Liu et al., 2021c), VL-T5 (Cho et al., 2021), VL-BART (Cho et al., 2021). However, for its capabilities of zero-shot predictions, we focused on the CLIP (Radford et al., 2021a) model for this dissertation. This model shows great promise in generating text in accordance with a given image. Furthermore, this model fits in the type of architecture we propose to prove as one of the most promising at this instant.

Secondly, we studied numerous models that propose to tackle automatic report generation based on radiology reports. However, most architectures are prepared to tackle this task in a more language-processing manner, having a lack of focus on clinical factuality. With models such as CXR-RePaiR (Endo et al., 2021), PPKED (Liu et al., 2021b), Clinical Transformer (Lovelace and Mortazavi, 2020), or even MDT + WCL (Yan et al., 2021), we have models that are aligned with the objectives of this dissertation. This means that all these models share in their proposal objectives such as increasing clinical fluency and accuracy, having some sort of memory retrieval system that enhances the generation process, and finally increasing the natural language processing capabilities. Although Transformer w/ RM + MLCN (Vaswani et al., 2017) was introduced in 2017, it is our understanding that this architecture is still very relevant in present studies, where we still see the MLCN (memory-driven conditional layer normalization) module present in works such as MDT + WCL (Yan et al., 2021), which dates to 2021.

Finally, we focus on models that have retrieval mechanisms to enhance their capacities (e.g. EXTRA (submission, 2022)). Although these models are very relevant to our work, we approach their study in order to determine the impact we can expect from such a methodology.



# 4

## Methodology

### Contents

---

4.1 Encoder-Decoder Architecture . . . . .	39
4.2 Retrieval Mechanism . . . . .	41
4.3 CLIP V&T . . . . .	42

---



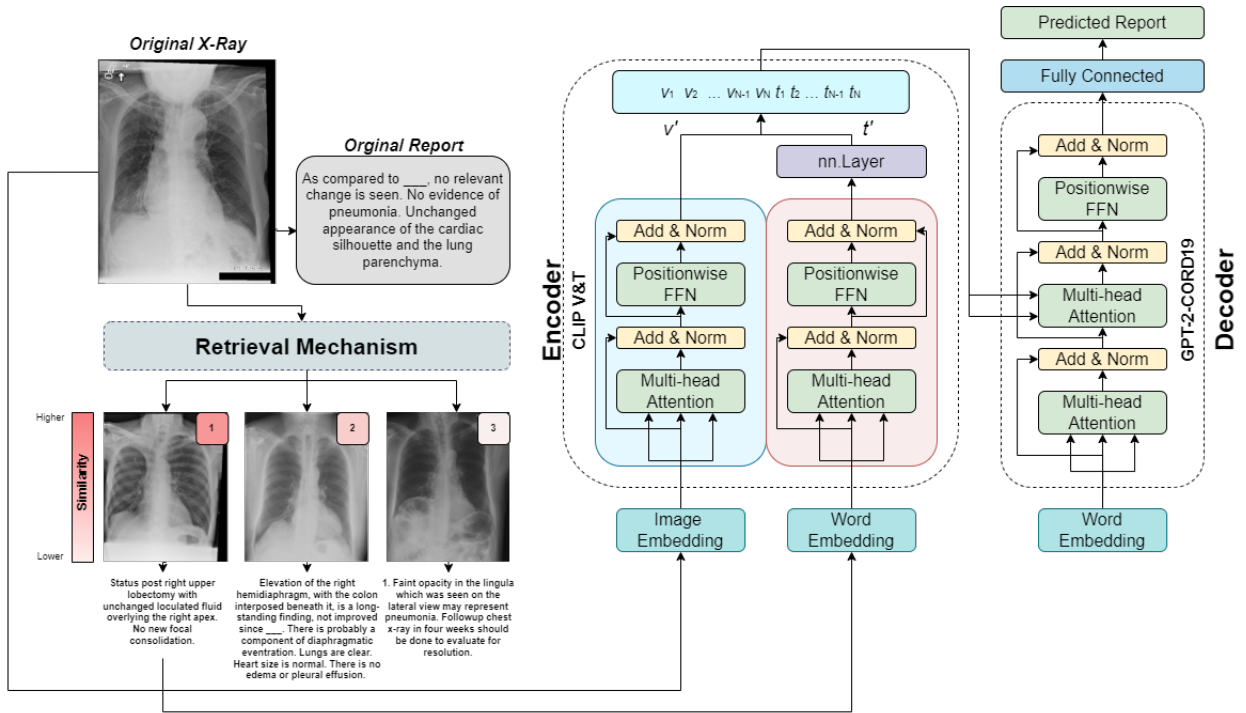


Figure 4.1: Architecture of the CLIP VT Model with retrieval mechanism.

Very similarly to works like PPKED (Liu et al., 2021b), CXR-RePaiR (Endo et al., 2021), or Clinical Transformer (Lovelace and Mortazavi, 2020), our proposed model is based on transformer introduced by OpenAI, called CLIP (Radford et al., 2021a). However, we use a pre-trained version of this transformer, where the data to which it was trained is also very similar to the MIMIC-CXR dataset. The model in question is MedCLIP<sup>1</sup>, where the encoder is the CLIP ViT-B/32 model (Dosovitskiy et al., 2020) and the decoder is BERT (Devlin et al., 2018). However, changes are made so that this transformer uses GPT-2 (Radford et al., 2019) as decoder, and later a pre-trained of GPT-2 on medical data, GPT-2-CORD19<sup>2</sup>. In this section, we will introduce our proposed model, as well as all the adaptations to make it final (i.e. the image-text encoder and retrieval mechanism).

## 4.1 Encoder-Decoder Architecture

The two architectures proposed in the current work, follow an encoder-decoder typology, in which it is employed the CLIP model Radford et al. (2021a), more specifically the MedCLIP model. In the first instance, the baseline will only be composed with an image encoder and text decoder, following the exact same architecture as in Image 2.5, and in a more advanced phase, we will make an association

<sup>1</sup><https://huggingface.co/flax-community/medclip>

<sup>2</sup><https://huggingface.co/mrm8488/GPT-2-finetuned-CORD19>

between image and text in the encoding process. Firstly, instantiated will be a baseline that will set ground scores to later determine the possible improvements of an enhanced version implementing an encoder capable of linking vision and language. This baseline will have a CLIP vision encoder (CLIP ViT-B/32 model (Dosovitskiy et al., 2020)), receiving a radiology image, and a CLIP text decoder (ClipBERT (Devlin et al., 2018)), that is set to receive the report according to the radiology image given to the encoder.

As mentioned before, this approach is based on the MedCLIP-roco, which is trained on the medical ROCO dataset Pelka et al. (2018). Following the previous approach, it is proposed to enhance the baseline version with an augmented retrieval mechanism, by which we adjust with more detail the best report to follow the radiology image in the encoding process. Consequently, this encoder is set to be enhanced itself, and for that, it is changed to not only accept images as inputs, but also the reports from the retrieval phase. This will result in an encoder that is both an image and text encoder. This doesn't apply any changes to the decoder, keeping the same decoder. However, the decoding algorithm is changed between greedy search to beam search (where the number of beams being  $b$ , varies in  $b = \{3, 4, 5\}$ ).

To encapsulate the models we propose to compare, we have what we designate as Baseline, where we only have the pre-trained encoder from MedCLIP while empowering it with GPT-2 (Radford et al., 2019). Secondly, we propose a model called Baseline w/ GPT-2-CORD19 changes to this baseline, so it can be more accurate in clinical analysis, changing the GPT-2 by a pre-trained version of it on medical data concerning COVID-19 data. Finally, we further propose a model that employs an image-text encoder composed by the CLIP ViT-B/32 model (Dosovitskiy et al., 2020) and CLIPBERT (Devlin et al., 2018), keeping GPT-2-CORD19 as the decoder. Furthermore, this model we call CLIP VT w/ Retrieval has a retrieval mechanism to further improve the generation of clinical reports.

#### 4.1.1 Encoder

The encoder presented in the current work has two main constructions. On the first set of tests to validate the capacity of the Transformer, we employ a visual encoder based on the CLIP ViT architecture. As results are promising, we propose the use of an encoder that can both deal with image and text inputs, later concatenating both representations into one, with non-changing dimensionality. The encoder, for the baseline model, is based on a pre-trained version of the CLIP ViT-B/32 model (Dosovitskiy et al., 2020) as mentioned before. This encoder takes a one-dimensional sequence of token embeddings generated from the pixel values of the input image. Using a constant vector size of  $D$ , in order to flatten the patches, so they can be mapped into the  $D$  dimensions, referred to as the patch embeddings. The encoder is prepared to retrieve features from images that are 268 by 268 pixels. This encoder is used in the MedCLIP-roco Radford et al. (2021a), being trained on the ROCO dataset (Pelka et al., 2018), with



81,825 radiology images. Before being trained on the ROCO, this encoder (Dosovitskiy et al., 2020) has been pre-trained on a dataset of 400 million image-text pairs.

### 4.1.2 Decoder

The decoder used in both models is a GPT-2 based (Radford et al., 2019) language model. In this model, we employ the cross-attention layer, in order to retrieve information directly from the encoder. This GPT-2-CORD19 encoder, enhanced with a pre-training phase on the CORD-19 dataset, exploits the capacity for suiting the generation on a more clinical level. This can provide the whole model the ability to better represent the radiology images fed to the encoder. There are also, within the decoder, masked multi-head self-attention layers, to avoid attending to tokens that may affect the next decoding phase. Also, still, on the cross-attention layer, this is employed so that we can assess the encoder's outputs and add the weights to the decoding phase, creating a link between the image and text. In a later instance, creating the link between the representation of both image and text, and the original report fed to the decoder. The generation will be done by predicting the next token, attending to the previous ones, and the encoder output. Finally, GPT-2 uses cross-entropy loss, or logarithmic loss, to measure the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverge from the actual label. So predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high loss value. A perfect model would have a logarithmic loss of 0.

## 4.2 Retrieval Mechanism

The retrieval mechanism employed to empower the generation is based on the Facebook AI Similarity Search (FAISS), where giving the vectors that represent any given data source, the  $k$  nearest neighbors are calculated using the Euclidean distance ( $L^2$ ). FAISS is also capable of doing so in an amount of time that does not worsen the complexity of the model. For this, there is the capacity of training and index each of these vectors to a data structure, where then it can be searched. Given a set of vectors  $x_i$  in dimension  $d$ , FAISS builds a data structure in RAM from it. After the structure is constructed, when given a new vector  $x$  in dimension  $d$  it performs efficiently the operation  $j = \operatorname{argmin}_i \|x - x_i\|$  where  $\|\cdot\|$  is the Euclidean distance. For many index types, this is faster than searching one vector after another to trade precision for speed, ie. giving an incorrect result 10% of the time with a method that's 10x faster or uses 10x less memory. The main structure used in FAISS, for this project, is a simple object file where we store all images in vectorial form, retrieved from the CLIP Radford et al. (2021a) model, making it easier to get the features of all images for later comparison by the retrieval mechanism. In this case, for any given image  $R$ , the retrieval mechanism will run on the given data structure with all vectors, to find the  $k$

nearest neighbors. Having the collection of  $k$  nearest elements, we then proceed to use the one that is closest to reality, in order not to negatively impact the generation and finally encode both the image and text.

### 4.3 CLIP V&T

To instantiate the enhanced version of this baseline, in order to receive both image and text as input, the encoder is expanded to be also a text encoder. Maintaining the CLIP Vision model (CLIP ViT-B/32), we also use the BERT text encoder used on the MedCLIP-roco [Radford et al. \(2021a\)](#), as it also is trained on the same medical data that the vision encoder already was trained. Given this, we then pass both image and text to what we call a vision-text encoder, working parallel on both inputs, where finally, given that  $b_s$  is the batch size and  $m_s$  is the maximum length for a sequence, the pooled output of shape  $\{b_s, 768\}$  with representations for the entire input sequences and a sequence output of shape  $\{b_s, m_s, 768\}$  with representations for each input token (in context). Then both  $v'$  and  $l'$ , denominating the pooled outputs for both the image and text encoder respectively, are concatenated, maintaining dimensionality by passing the pooled output of the text encoder to a linear layer, resulting on a single representation for the image-text pair  $E_o = \{v_1, v_2, \dots, v_n, l_1, l_2, \dots, l_m\}$ , where  $E_o$  is the encoder output. Furthermore, the nn.Layer is crucial to keep the dimensions of the text encoder with the image encoder, as text representations dimensions are variable since clinical reports do not respect a word count or limit. Finally, This joint pooled output is then used to generate the attention that will represent uniquely what should be the visual-text encoder attention over both inputs, to be then passed to the decoder.

# 5

## Experimental Evaluation

### Contents

---

5.1 Experimental Setup . . . . .	45
5.2 Experimental Results . . . . .	47

---



Greedy Search				
Model with Greedy	BL-1	BL-4	MTR	RG-L
<b>(Ours) CLIP vision and text.</b>	26.6	9.8	23.9	35.9
Beam Search				
beams (b)	BL-1	BL-4	MTR	RG-L
b = 3	40.19	11.6	26.9	38.6
b = 4	40.16	11.9	27.04	39.2
b = 5	40.2	12.1	27.7	40.3

**Table 5.1:** Beam Search decoding method evaluation on the CLIP vision and text model.

## 5.1 Experimental Setup

To elevate the capacity of making the proposed model perform better in the radiology report creation, there was an increased concern to adjust both datasets and evaluation metrics to the extent of being in the same level of commitment as the SOTA models presented throughout.

### 5.1.1 Dataset

To evaluate both models presented, we opt to use the MIMIC-CXR dataset (Johnson et al., 2019), a large dataset of medical radiology studies. Each study contains a pair of x-ray images and the report for that same image. The organized splits were applied according to the specifications of the dataset. For that, from a total of  $\approx 85k$  studies,  $\approx 85\%$  for the training set, and  $\approx 7.5\%$  for both validation and test sets. Thus, this gives 65,567 studies for the training set and 5,000 for the remaining test and validation.

### 5.1.2 Evaluation Metrics

In order to quantify the quality of the generation of both models, we propose to use metrics that will represent coherence and factuality. For that metrics such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for a more detailed assessment of the two approaches taken.

### 5.1.3 Hyperparameters Fine-tuning

To understand how much better the proposed Model would behave, there were some fine-tuning tests to rectify misleading behavior. This means there are cases where the model simply does not perform as it is expected, creating too many mistakes in the generation process, or falling short against outgrown models.

To minimize the loss function and in order to update the weights in each epoch, the AdamW optimizer function was used, which by itself it is the improved version of the Adam (Loshchilov and Hutter, 2017) optimizer. This function takes into account a learning rate  $lr$ , coefficients used for computing running averages of gradient and its square  $\beta_1$  and  $\beta_2$ , a term  $\epsilon$  to be added to the denominator in order to improve numerical stability, the weight decay coefficient  $\lambda$ , and finally the option of using the AMSGrad (Reddi et al., 2018) variant of the Adam algorithm. For the learning rate, several values were tested,  $lr = \{1e-3, 2e-3, 3e-3\}$ , where results back a learning rate of  $1e-3$ . Concerning the betas and the term added to the denominator, they were kept as default as  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 1e-8$ . Following the work done on Adam (Loshchilov and Hutter, 2017), the best weight decay coefficient  $\lambda$  was appointed as 0.02, showing the best results, and that is the one used.

Following this process, there was also proposed to test several decoding methodologies, to assess the change in performance from one to another. As it is shown in the 5.1, the decoder was tested as a greedy decoder against a Beam Search decoder with different beam sizes  $b = \{3, 4, 5\}$ . It is important to indicate that the model used in this phase was the Clip vision and text, which is the improved model for this Thesis work. Assuming that the environment of the test is the same throughout, there is a clear indication that using a Beam Search decoder will improve the model's performance according to every score metric used. For this reason, the model uses a Beam Search decoder with 5 beams.

#### 5.1.4 Evaluation Methodology and Training

Both models implement a similar strategy, using an encoder-decoder typology. However, the Baseline model will employ a CLIP-ViT-B32 encoder, with a GPT-2-finetuned-CORD19 text decoder, which is based on the original GPT-2 but trained on the CORD-19 dataset improving the generation of medically driven sentences. The GPT-2 is a twelve-layer decoder-only transformer, using twelve masked self-attention heads, with 64-dimensional states each (for a total of 768). Although the decoder remains the same, on the enhanced version of this model we propose to use both text and vision combined in a single encoder. Consequently, we use the same vision encoder CLIP-ViT-B32 and BERT as the text encoder, with 12 encoders with 12 bidirectional self-attention heads pre-trained from unlabelled data extracted from the BooksCorpus with 800M words and English Wikipedia with 2,500M words. Following this, the use of the library FAISS in order to set up the retrieval mechanism, leveraging GpuIndexIVFFlat trained on the 55,675 image vectors. To store all the image vectors for posterior retrieval, an object file is created containing all 55,675 radiology images with the vectorial representation given by the model CLIP-ViT-B32. Then, an L2-Similarity is calculated between the report and the input image, and also the input image and other images, always retrieving a set of  $k$  nearest neighbors. The value of  $k$  can be such that  $k = \{1, 2, 3\}$

## 5.2 Experimental Results

In this section we will explore the results of several experiments made using both models presented in this dissertation, focusing on the CLIP Image-Text encoder model and the use of the new methodologies we propose to enhance this model, and its performance over the state-of-the-art. We further study the behavior of the most promising model (CLIP VT) on the generation of clinical reports, in the clinical and natural language processing sense, while verifying how the attention in the encoder behaves when parsing the x-ray. We want to answer questions such as:

- How do both models (Baseline and CLIP Vision-Text) perform against state-of-the-art models?
- Will the vision-text encoder outperform normal vision encoders?
- How much can we increase performance by using a retrieval mechanism?
- How accurate and eligible (in clinical terms) are the reports created?
- Where are the most attended spots by the model in the x-ray?

### 5.2.1 Report Generation Models Performance

Over the past few years, models focused on generating reports according to the analysis of x-rays have increased at an interesting pace. Most of the models we compare our work to have provided ground to improve upon such tasks.

In 5.2 there are present some of the most predominant models used for report generation over the use of x-rays. On top, we present two models that are based on models other than Transformers. Those are the CNN + RNN and LSTM with CoAttention, both tested on IU-XRay. This dataset is very similar to MIMIC-CXR, for which we can evaluate the performance of both given the similarity. According to the other models present in the same table, those are trained and tested on the same dataset, and given that, we can more accurately conclude if our models perform close or better compared to these state-of-the-art models.

From 5.2 we introduce a clear comparison from the standpoint of performance, in what concerns models with similar mechanisms compared to our CLIP Vision and Text model. Although encoders and decoders in these models may vary, most models also provide a technique proposing the review of x-rays before introducing the original to the generation.

According to our results, we can see that upon the BLEU metric, our Baseline falls short performance-wise, compared to the most predominant methods (e.g. Nguyen et al. ([Nguyen et al., 2021a](#)) and Clinical Transformer ([Lovelace and Mortazavi, 2020](#))). Following this measure, we get adequacy and fluency, according to a score from 1 to 100. The closer to 100, the more adequate and fluent the generated text

Model	BL-1	BL-2	BL-4	C	M	RG-L
Clinical Transformer	41.5	27.2	14.6	-	15.9	31.8
Transformer w/ RM	32.4	19.6	9.5	-	12.8	26.5
Transformer w/ RM + MLCN	35.3	21.8	10.3	-	14.2	27.7
$M^2$ TR.	36.1	22.1	10.1	-	13.9	26.6
$M^2$ TR. Progressive	37.8	23.2	10.7	-	14.5	27.2
PPKED	36.0	22.4	10.6	23.7	14.9	28.4
Align Transformer	37.8	23.5	11.2	-	15.8	28.3
Nguyen et al.	49.5	36.0	22.4	-	22.2	39.0
MDT + WCL	49.5	36.0	22.4	-	22.2	39.0
<b>(Ours) Baseline</b>	32.6	22.6	10.2	27.2	23.8	35.5
<b>(Ours) Baseline w/ Cord19</b>	33.4	22.8	11.5	26.3	24.5	36.3
<b>(Ours) CLIP VT w/ Retrieval</b>	40.2	26.5	12.1	29.2	27.7	40.3

**Table 5.2:** NLP performance results containing the MIMIC-CXR and IU-Xray datasets.

is, and it means there starts to exist an overlap with human translation texts. Given this, we can see that our CLIP VT with retrieval provides a slight increase in the fluency presented on text, presenting results equivalent to well-founded models.

As compared within the same range, the CiDER metric provides an insight into the increase in performance we can achieve by simply fine-tuning the model and providing the mechanism of retrieval, as well as introducing a more adequate decoder (GPT-2 trained on Cord19). As we can see, the consensus-based image description evaluation metric will provide a result on the correlation between text and image. As we can see from the results, performance increased in every new model definition.

Interestingly, according to the METEOR metric, where the score is given according to the translation alignment. This metric is actually relevant to our work since it gives insight into the correlation between our models generated text and text written by Humans. The metric by itself has shown a correlation of 0.964 with human judgment at the corpus level, compared to 0.817 on the same data set. Given this, we can see that our most basic model (Baseline) performs better than any other model, with a difference of 1.6 points to Nguyen et al (Nguyen et al., 2021a). This is a great achievement on its own, supporting that every other enhancement can even provide better correlation results.

Finally, on ROUGE-L, the metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. Again, we can see that even our Baseline has very competitive results, lacking only 3.5 points from the most predominant model. However, when compared to our CLIP VT With retrieval, the increase in performance is noticeable.

## 5.2.2 Clinical Efficiency Evaluation

While most works focus on NLP metrics such as ROUGE-L or CiDER, these on their own do not explain how the models will behave when it comes to clinical accuracy. The main focus should be to get a



Model	P	R	F1
$M^2$ TR.	32.4	24.1	27.6
$M^2$ TR. Progressive	24.0	42.8	30.8
MDT + WCL	38.4	27.4	29.4
Clinical Transformer	41.1	47.5	36.1
<b>(Ours) CLIP VT w/ Retrieval</b>	28.4	34.7	31.2

**Table 5.3:** Clinical accuracy results on some of the Transformer Models.

model as fluent and semantically correct as possible, while also achieving good clinical accuracy. This imbalance, in some cases, will make a model good for generating text, however, unusable when it comes to the medical purposes for which it stands.

To provide a comprehensive overview of the behavior of some models in what concerns clinical accuracy, we have selected such as presented in 5.3. These transformer-based models achieve great NLP results. However, let us take by example the  $M^2$  TR. Progressive model, achieving results such as 27.2 on ROUGLE-L, but when it comes to precision according to the medical features in the report, it achieves 30.8 in the F1 metric. This means that the overall accuracy of the  $M^2$  TR. Progressive model over the MIMIC-CXR dataset is low. Following this conclusion, our model () also presents results that are subpar to what we hoped to achieve. Nonetheless, compared to the state-of-the-art it keeps presenting competitive prospects.

### 5.2.3 Retrieval Mechanism Evaluation

Although we have seen that our latest model outperforms others in most metrics, we have to test the capacity of this retrieval mechanism, for further betterment of the CLIP VT model. Consequently, we have proposed a series of tests concerning only the retrieval mechanism. Firstly, we propose the evaluation of the performance of this mechanism from two different perspectives, finding the more similar image and retrieving the report given that same image, and finally retrieving the report from the closest report given the query image. Following the conclusions of this evaluation on 4, we vary the k, given that k is the number of neighbor images on the cluster. Furthermore, the relevance of these results will directly impact the model by itself, since if we increase the performance of the retrieval mechanism the better we prepare the model for the generation of even more accurate and fluent reports, as well as increasing the accuracy of the medical component.

As depicted on 5.6, we provide an insight into whether retrieve the reports given the closest image or the closest report. From the standpoint of performance, we can clearly see that by finding the reports given the closest image to the x-ray query, we can have a better correlation with human texts, as well as fluency. These results are not short of logical if we think about the X-ray structure. Images are very similar to one another, although, the details (white areas of the x-ray) are where the analysis has

k	BL-1	BL-4	MTR	RG-L
k = 1	37.3	10.1	24.5	38.1
k = 2	39.8	10.4	26.2	39.7
k = 3	40.2	12.1	27.7	40.3

**Table 5.4:** Retrieval mechanism evaluation (of CLIP VT w/ Retrieval) on the capacity of retrieving reports with close meaning to the original report, regarding image-to-image similarity, with k neighbors variation.

Technique	BL-1	BL-4	MTR	RG-L
Similar Report	39.1	11.7	25.2	37.6
Similar Image	40.2	12.1	27.7	40.3

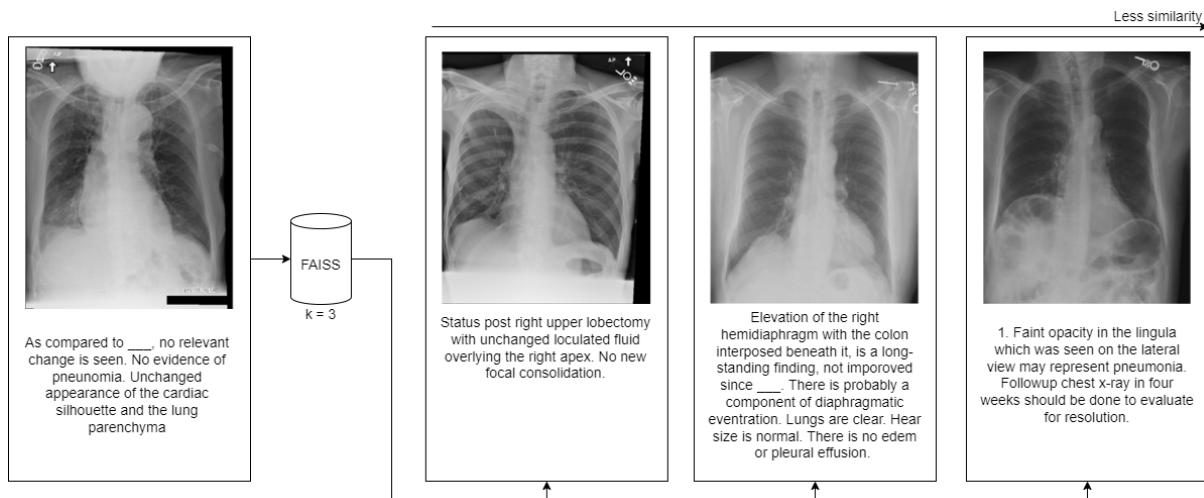
**Table 5.5:** Retrieval mechanism evaluation (of CLIP VT w/ Retrieval) on image-to-image similarity compared to image-to-text similarity.

Original	Retrieved
A comparison with the study of , there is no change or evidence of acute cardiopulmonary disease. No pneumonia, vascular congestion, or pleural effusion.	Since the prior study obtained the same early there is no change in the position of the Port-A-Cath catheter as well as multiple metastatic nodules projecting over the chest bilaterally. No evidence of subcutaneous air was demonstrated on the current examination.
No acute cardiopulmonary process. No pneumothorax.	Comparison to . No relevant change is noted. No pneumonia, no pulmonary edema, no pleural effusions. Normal size of the heart.
No acute findings.	No acute cardiopulmonary abnormality.

**Table 5.6:** Comparison between some results of the retrieval mechanism.

differed. Although it might seem that two random images are close to the naked eye, they may be very distinct according to the vision encoder feature extraction. Furthermore, to provide more insight into this argument, the deflation of a right lung, compared to the normal capacity of another right lung is very similar to our eye, but not when the image is represented in vectorial form, where 0's are 1's in the other picture.

On 5.4, and following past results, we vary the number of reports that might be retrieved from the similarity process. Simply, where we used to retrieve only 1 report, we can now retrieve k reports from the k nearest images to the query x-ray. Although the closest image can present a good report, there can be other images, also similar to the query, that offer a more structured report with more detailed information. Following this statement, we can see that we achieve better results when we provide a larger set of reports to which we can retrieve larger and better features from the text.



**Figure 5.1:** Retrieval process for  $k = 3$  nearest neighbors and respective reports.

## 5.2.4 Report Generation Study

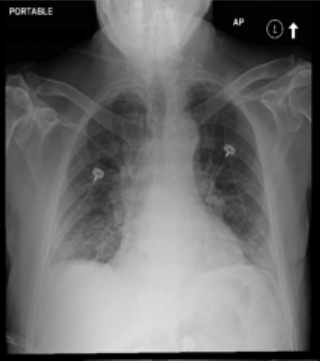
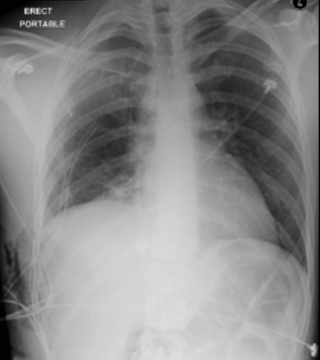

This study was prepared so that we can visualize and assess the capacity of generating accurate medical reports, as well as semantically accurate reports. The importance of this evaluation lies since we want to achieve the best reports on both semantics and medical notions.

As depicted in 5.2, we took three randomly chosen x-rays and fed the model in order to generate the reports for each one of those. In this figure, we can see the x-ray and the report that pairs with it. To this report, we denominate as Actual, and as for the generated report, we state it as Prediction.

On the top image, the baseline report appears with a more dense structure, proposing that more information might be present, in a first glance when compared to the prediction. However, if we analyse both reports side by side, we can see that the prediction ascertains over 83% of the actual medical indications, missing details such as indicating the presence of catheters in both lungs, not just in the left lung, as indicated in the prediction.

According to the image in the middle, we can state that the report is a brief summary of the actual report, making the medical analysis as brief as possible. This result is remarkable as we achieve 100% of medical accuracy based on the ground truth, as well as eliminating information that does not add any relevance to the report.

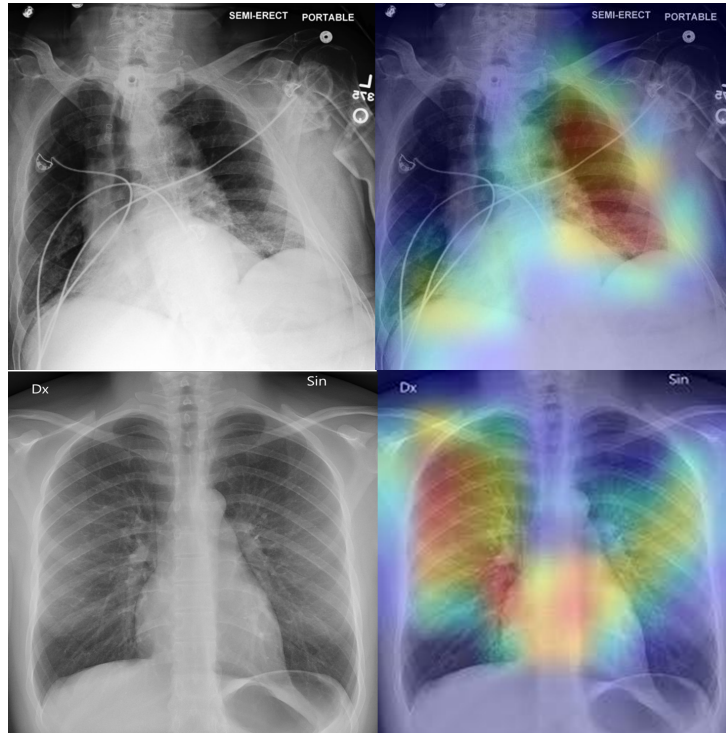
Finally, the bottom x-ray report is very brief, contrary to the prediction. For this, we actually achieve the opposite result as stated in the last paragraph. The model tried to indicate medical features that were not initially in the actual report. However, not all added information is inaccurate. For instance, the indication of no pneumothorax is correct, as we don't see any deflation of any lungs. Also, the correct evidence of no pneumonia is due to the clarity of the lungs.

	<b>Actual</b>	In comparison to ___ chest radiograph, lung volumes are lower, accentuating the cardiac silhouette and resulting in crowding of bronchovascular structures. New patchy bibasilar opacities could be secondary to aspiration, developing infectious pneumonia and atelectasis. Exam is otherwise remarkable for an apparent new small right pleural effusion.
	<b>Prediction</b>	Mild - sided basal atelectasis. Mild cardiomegaly. Mild left lower lobe atelectasis and small right middle lobe pneumonia and bilateral pleural effusion. Catheter in appearance on the left lung.
	<b>Actual</b>	2 right chest tubes are in place with subsequent interval decrease in the right pleural effusion. Heart size and mediastinum are stable. Left lung is clear. Subcutaneous air within the right chest wall is small
	<b>Prediction</b>	Normal heart size. No acute cardiomegaly. Conclusion : Right pleural effusion.
	<b>Actual</b>	Unchanged bilateral areas of opacification and consolidation.
	<b>Prediction</b>	Mild cardiomegaly has cleared. There are lower omediastinal lung volumes catheter. No pneumonia. No new opacification within the bilateral areas. No abnormality or pneumothorax.

**Figure 5.2:** Three randomly chosen results of the CLIP VT w/ Retrieval model, showing the practical results on the generation of new clinical reports.

## 5.2.5 Contrastive Attention Study

The contrastive attention study is the last evaluation step of the model, where we will assess where the vision encoder is extracting the features, to later pass to the decoder where these features will have a meaning. Following this, results on the attention maps will ideally show more detail on the lung and heart areas, where most of the diagnosis is made. The bone structure should not be left out completely, but these are not the main focus of the reports, as we have come to state throughout the dissertation. In this segment of results, we will see attention maps on two x-rays from the MIMIC-CXR dataset as seen in 5.3, and those will lead to conclusions as, for example, if the model is looking to the right areas and



**Figure 5.3:** Contrastive attention study done on two randomly chosen x-rays from the MIMIC-CXR dataset.

retrieving features with relevance for the generation.

As seen from the image, the CLIP VT model has assured that most of the lung area and heart are covered with great detail and attention. The measurements indicate that, if the color in the gradient is closer to red, the attention scores are higher, meaning that the model has focused on those areas with greater attention.

The results on these two x-rays, where the model focuses on the lungs and heart area most of all, leave us to conclude that the reading of the image features is being done correctly, or as close to expected as we wanted.

## 5.2.6 Overview

This chapter encapsulates all the results of testing the models proposed. Starting with an overall NLP and clinical accuracy study of the CLIP-based models, we can clearly see that, however not being the best performing model, it reaches competitive results, leaving a margin for improvement. This is also complemented by the results present in the contrastive attention map of the x-ray in Figure 2.6, and then by the generation results in Figure 5.1 and 5.2. Furthermore, we achieve the best overall results in CIDEr, METEOR, and ROUGE-L, among models that are renowned in this scientific area.

Finally, regarding the capacity of the retrieval enhancement we propose, by the table of results 3.1,

we can see that in comparison with the starting models (Baseline and Baseline with GTP-2-CORD19), the information retrieval has a positive impact in the generation phase, where we further prove that for this specific case and architecture, we should retrieve reports regarding similarity as image-to-image, as the results in Table 5.6 shows.

# 6

## Conclusion

### Contents

---

6.1 Conclusions . . . . .	57
6.2 System Limitations and Future Work . . . . .	57

---





## 6.1 Conclusions

In this work, a CLIP Transformer is used to assess the improvements that can be achieved in the report generation of thorax x-rays. This is used as a retrieval-augmented report generator, and in some cases, it was shown to improve the performance while using retrieved medical images to indicate the best-suited report to follow generation. Also, the model exploits different encoding methods, where not only the image is the input, but both the image and report are the input for a CLIP Image-Text encoder. The encoding process by itself has shown to be of some significance as it concatenates both image and text into a single representation, that being already a proposition of the CLIP Model [Radford et al. \(2021a\)](#). By doing so, this representation elevates the capacity of creating unique representations. Upon the evaluation process on the MIMIC-CXR dataset, there is room to safely say that this model suggests some improvements in an overall case.

## 6.2 System Limitations and Future Work

For future work, this model can be utilized with some more details in mind. As discussed, the metric CLIPScore can enhance CLIP models to better adjust the image to a text representation. This is one point where CLIPScore might be used, as a metric for guiding the generation, over the representation of the encoder. Also, in the future, more than two datasets should be used in order to better validate the capacity of the model, and the mechanism proposed. For this work, we have limited the options of generation to MIMIC-CXR, where the IU-Xray could also be used in later proceedings. In the retrieval phase, there should not be full reliability only on the L2 similarity, but on other metrics as well, making the retrieved report more suitable. In the present case, some reports for some images are repeated and don't increase the value of the original report. Finally, some train should be done on the retrieval mechanism, to better prepare it for the medical imagery retrieval task.



# Bibliography

- Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Banerjee, S. and Lavie, A. (2005). 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*, pages 65–72.
- Boag, W., Hsu, T.-M. H., McDermott, M., Berner, G., Alesentzer, E., and Szolovits, P. (2020). Baselines for chest x-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140. PMLR.
- Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.
- Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B., Langlotz, C. P., Amrhein, T. J., and Lungren, M. P. (2018). Deep learning to classify radiology free-text reports. *Radiology*, 286(3):845–852.
- Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
- Cho, J., Lei, J., Tan, H., and Bansal, M. (2021). Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*.
- Cho, J., Yoon, S., Kale, A., Derroncourt, F., Bui, T., and Bansal, M. (2022). Fine-grained image captioning with clip reward.
- Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

- Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Hounsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.
- Endo, M., Krishnan, R., Krishna, V., Ng, A. Y., and Rajpurkar, P. (2021). Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR.
- Fei, Z. (2021). Memory-augmented image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1317–1324.
- Harzig, P., Chen, Y.-Y., Chen, F., and Lienhart, R. (2019). Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*.
- Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.
- Ippolito, D., Kriz, R., Kustikova, M., Sedoc, J., and Callison-Burch, C. (2019). Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362*.
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilicus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.
- Jing, B., Xie, P., and Xing, E. (2017a). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
- Jing, B., Xie, P., and Xing, E. (2017b). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

- Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with GPUs.
- Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2021). Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*.
- Larochelle, H. and Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order boltzmann machine. *Advances in neural information processing systems*, 23:1243–1251.
- Lee, H., Yoon, S., Dernoncourt, F., Bui, T., and Jung, K. (2021). Umic: An unreferenced metric for image captioning via contrastive learning. *arXiv preprint arXiv:2106.14019*.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., and Choi, Y. (2021a). Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
- Liu, F., Wu, X., Ge, S., Fan, W., and Zou, Y. (2021b). Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.

- Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.
- Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021c). Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*.
- Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization.
- Lovelace, J. and Mortazavi, B. (2020). Learning to generate clinically coherent chest x-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243.
- McDermott, M. B., Hsu, T. M. H., Weng, W.-H., Ghassemi, M., and Szolovits, P. (2020). Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR.
- Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P., and Jurafsky, D. (2020). Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*.
- Mokady, R., Hertz, A., and Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.
- Najdenkoska, I., Zhen, X., Worring, M., and Shao, L. (2021). Variational topic inference for chest x-ray report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 625–635. Springer.
- Nguyen, H. T., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., and Cheng, L. (2021a). Automated generation of accurate & fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*.
- Nguyen, H. T., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., and Cheng, L. (2021b). Automated generation of accurate & fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*.
- Nooralahzadeh, F., Gonzalez, N. P., Frauenfelder, T., Fujimoto, K., and Krauthammer, M. (2021). Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Parvaiz, A., Khalid, M. A., Zafar, R., Ameer, H., Ali, M., and Fraz, M. M. (2022). Vision transformers in medical computer vision – a contemplative retrospection.

- Pelka, O., Koitka, S., Rückert, J., Nensa, F., and Friedrich, C. M. (2018). Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189. Springer.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021a). Learning transferable visual models from natural language supervision.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021b). Learning transferable visual models from natural language supervision.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. In *International Conference on Learning Representations*.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- Sarto, S., Cornia, M., Baraldi, L., and Cucchiara, R. (2022). Retrieval-augmented transformer for image captioning. *arXiv preprint arXiv:2207.13162*.
- Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., and Langs, G. (2015). Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer.
- Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.
- Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. (2022). Transformers in medical imaging: A survey.
- Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2021). How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.
- Singla, K., Pressel, D., Price, R., Chinnari, B. S., Kim, Y.-J., and Bangalore, S. (2022). Cross-stitched multi-modal encoders.

- Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., and Lungren, M. P. (2020). Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.
- submission, A. A. (2022). Retrieval-augmented image captioning.
- Syeda-Mahmood, T., Wong, K. C., Gur, Y., Wu, J. T., Jadhav, A., Kashyap, S., Karargyris, A., Pillai, A., Sharma, A., Syed, A. B., et al. (2020). Chest x-ray report generation through fine-grained label learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 561–571. Springer.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator.
- Volodina, O. V. and <https://pnojurnal.wordpress.com/2022/07/01/volodina-3/> (2022). Formation of future teachers' worldview culture by means of foreign-language education. *P Sci Edu*, 57(3):126–159.
- Wang, L., Bai, Z., Zhang, Y., and Lu, H. (2020). Show, recall, and tell: Image captioning with recall mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12176–12183.
- Xu, C., Yang, M., Ao, X., Shen, Y., Xu, R., and Tian, J. (2021). Retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning. *Knowledge-Based Systems*, 214:106730.
- Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., and Hsu, C.-N. (2021). Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*.
- You, D., Liu, F., Ge, S., Xie, X., Zhang, J., and Wu, X. (2021). Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–82. Springer.
- You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.



Zarrieß, S., Voigt, H., and Schüz, S. (2021). Decoding methods in neural language generation: A survey. *Information*, 12(9):355.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2021). Scaling vision transformers. *arXiv preprint arXiv:2106.04560*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.