Neural Models for Generating Clinically Accurate Chest X-Ray Reports

André Loras Leite*
Department of Computer Science and Engineering
Instituto Superior Técnico
Lisbon, 1049-001, PT
andre.l.leite@tecnico.ulisboa.pt

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 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, and 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.

1 Introduction

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.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

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.

2 Related Work

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.

2.1 General Image Captioning

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.

2.2 Report Generation

Numerous models 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), Align Transformer (You et al., 2021), the proposal by Nguyen et al. (2021a), M^2 TR. Progressive or M^2 TR. (Nooralahzadeh et al., 2021), 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.

2.3 Retrieval Mechanism in Image Captioning Tasks

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

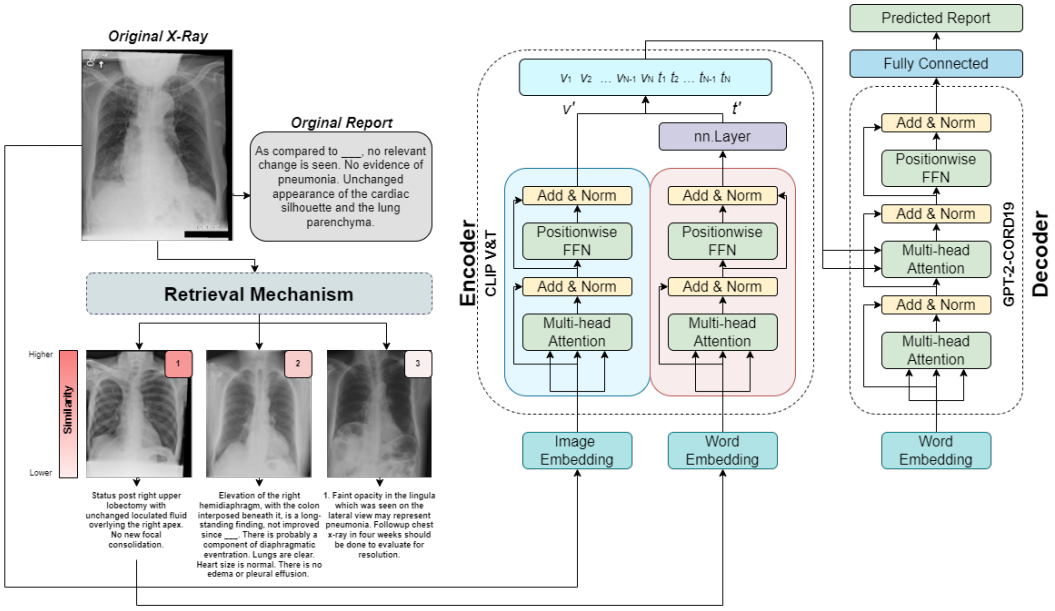

Figure 1: Architecture of CLIP VT with the retrieval mechanism.

in order to determine the impact we can expect from such a methodology. Works like the proposal Fei (2021) and (Sarto et al., 2022), have complemented models of machine learning (more concretely Transformers), with the capacity of retrieving already known information, in order to better adjust the generation given information that might be lacking from the original data. A good example is CXR-RePaiR (Endo et al., 2021), where they propose to enhance a model for generating x-rays, with the retrieval mechanism enhancement, being very aligned with the work we are presenting.

3 Methodology

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[2], 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[3]. 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).

3.1 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 **??**, and in a more advanced phase, we will make an association 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.

[2]https://huggingface.co/flax-community/medclip
[3]https://huggingface.co/mrm8488/GPT-2-finetuned-CORD19

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.

3.2 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.

3.3 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.

3.4 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 $j = argmin_i \| x - x_i \|$ 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.

3.5 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 $E_o = \{ v_1, v_2, ..., v_n, l_1, l_2, ..., l_m \}$, 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.

4 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.

4.1 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.

4.2 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.

4.3 Hyperparameters Fine-tuning

To understand how much better the proposed Model would behave, there were some fine-tuning tests to rectify misleading behaviour. By this, meaning cases where the model simply does not perform as

Greedy Search				
Model with Greedy	BL-1	BL-4	MTR	RG-L
(Ours) CLIP vision and text.	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	45.6
b = 4	40.16	11.9	27.04	46.2
b = 5	40.2	12.1	27.7	46.3

Table 1: Beam Search decoding method evaluation on the CLIP vision and text model.

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 ϵ to be added to the denominator in order to improve numerical stability, the weight decay coefficient λ, and finally the option of using the AMSGrad (Reddi et al., 2018) variant of the Adam algorithm. For the leaning 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 λ 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 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 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.

4.4 Evaluation Methodology and Training

Both models implement a similar strategy, using an encoder-decoder typology. However, in 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 medical 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). Altough 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 vectorial representation given by the model CLIP-ViT-B32. Then, a L2-Similiarity is calculated between the report and the input image, and also the input image and other images, always retrieving a set of k nearest neighbours. The value of k can be such that $k = \{\, 1, 2, 3 \,\}$

5 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.

5.1 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 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 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. Ngyuen 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 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 Ngyuen 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 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 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 3. These transformer-based models achieve great NLP results. However, let us take by example the M^2 TR. Progressive model, achieving results such

Model	Dataset	BL-1	BL-2	BL-4	C	M	RG-L
CNN + RNN	IU-XRay	-	-	9.5	11.1	15.9	26.7
LSTM w/ CoAttention	IU-XRay	-	-	24.7	32.7	21.7	44.7
Clinical Transformer	MIMIC-CXR	41.5	27.2	14.6	-	15.9	31.8
Transformer w/ RM	MIMIC-CXR	32.4	19.6	9.5	-	12.8	26.5
Transformer w/ RM + MLCN	MIMIC-CXR	35.3	21.8	10.3	-	14.2	27.7
M^2 TR.	MIMIC-CXR	36.1	22.1	10.1	-	13.9	26.6
M^2 TR. Progressive	MIMIC-CXR	37.8	23.2	10.7	-	14.5	27.2
PPKED	MIMIC-CXR	36.0	22.4	10.6	23.7	14.9	28.4
Align Transformer	MIMIC-CXR	37.8	23.5	11.2	-	15.8	28.3
Ngyuen et al.	MIMIC-CXR	49.5	36.0	22.4	-	22.2	39.0
MDT + WCL	MIMIC-CXR	49.5	36.0	22.4	-	22.2	39.0
(Ours) Baseline	MIMIC-CXR	32.6	22.6	10.2	27.2	23.8	35.5
(Ours) Baseline w/ Cord19	MIMIC-CXR	33.4	22.8	11.5	26.3	24.5	36.3
(Ours) CLIP VT w/ Retrieval	MIMIC-CXR	40.2	26.5	12.1	29.2	27.7	40.3

Table 2: Results containing the MIMIC-CXR and IU-Xray datasets.

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

Table 3: Clinical accuracy results on some of the Transformer Models.

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.

5.3 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 **??**, 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, 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 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.

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

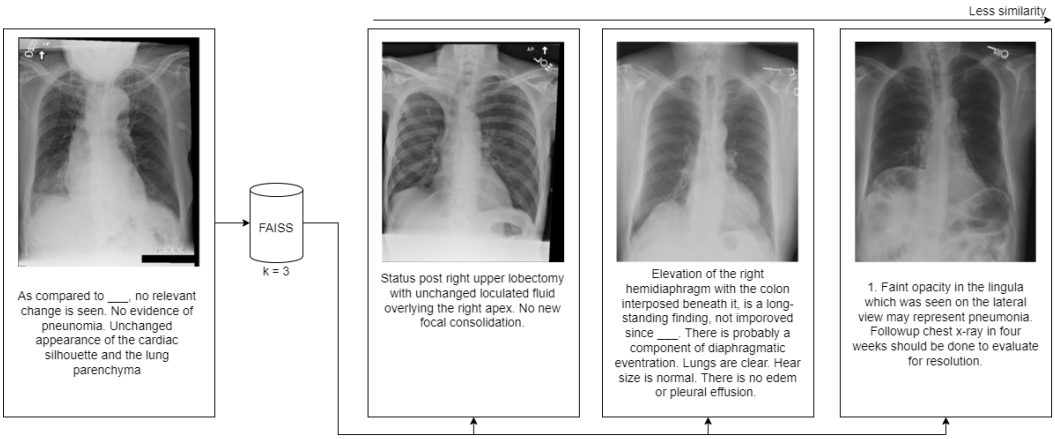

Figure 2: Retrieval process for k = 3 nearest neighbors and respective reports.

On 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.

5.4 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 3, 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.

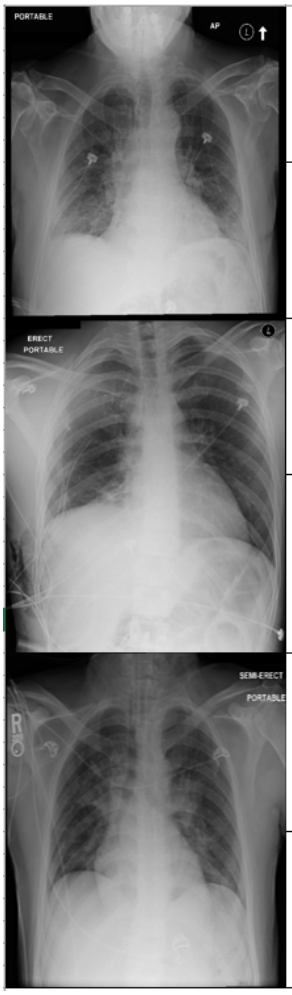	**Actual**	In comparison to ___ chest radiograph, lung volumes are lower, accentuating the cardiac silhouette and resulting in crowding of bronchovascular structures. New patchy bibasilar opacities could be secondary to aspiration, developing infectious pneumonia and atelectasis. Exam is otherwise remarkable for an apparent new small right pleural effusion.
	Prediction	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.
	Actual	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
	Prediction	Normal heart size. No acute cardiomegaly. Conclusion : Right pleural effusion.
	Actual	Unchanged bilateral areas of opacification and consolidation.
	Prediction	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 3: Retrieval mechanism evaluation (of CLIP VT w/ Retrieval) on image-to-image similarity compared to image-to-text similarity.

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.

5.5 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 4, and those will lead to conclusions as, for example, if the model is looking to the right areas and retrieving features with relevance for the generation.

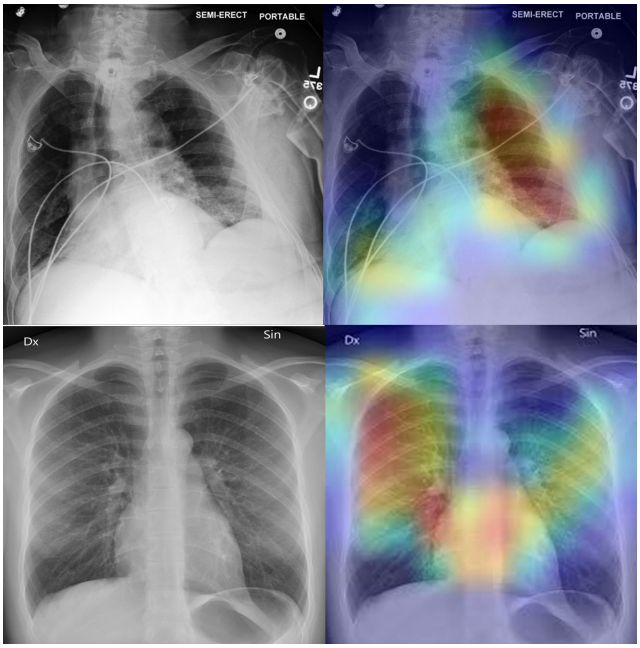

Figure 4: Contrastive attention study done on two randomly chosen x-rays from the MIMIC-CXR dataset.

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.

6 Conclusions and Future Work

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.

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.

References

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., Dernoncourt, 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 Houlsby, 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-Ilcus, S., Chute, C., Marklund, H., Haghgoo, 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://pnojournal.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*.