Clinical Report Generation Powered by Machine Learning

Iulia-Renata Sîrbu
ZHAW School of Engineering
Master of Data Science
Exchange Student 2022-2023

sirbuiul@zhaw.students.ch
iulia.renata.sirbu@gmail.com

Dr. Jasmina Bogojeska
ZHAW School of Engineering
Explainable Artificial Intelligence Group
VT Coordinator Spring 2023

jasmina.bogojeska@zhaw.ch

Abstract

Medical imaging is crucial for diagnosing, monitoring and treating medical conditions. Their corresponding medical reports are the primary medium through which medical professionals attest their findings. The automated generation of radiography reports has thus the potential to improve and standardize patient care and significantly reduce clinicians workload. Through this project, we have researched existing methods in machine learning for the task of automated image captioning as well as tested different approaches using only transformer based methods in order to generate radiology reports for medical images. The experiments have been conducted using the MIMIC-CXR database, and the results obtained are comparable with current state-of-the-art on both natural language generation metrics including RougeL, Meteor and classification metrics such as F1 macro and F1 micro that were used in order to assess the clinical accuracy and completeness of the proposed method.

1 Introduction

A. Context The task of image captioning is a highly researched one [1, 2]. It yields many benefits in various fields. In the field of medicine, image captioning could help clinicians make more accurate and faster diagnoses. After the assessment of the medical images, in this case, the radiography images, the clinicians state their findings and diagnoses through written reports. The information the clinicians provide in the image reports is highly valuable for many reasons: for the future assessment of the patient so as to monitor the development of the health condition or for a second opinion of another doctor.

B. Problem The writing of the report is a costly procedure that takes much time and expertise on the account of the medical professional assessing the patient's health condition. This cost can be alleviated with the use of machine learning models that could perform the image captioning task before the clinician, reducing the time necessary for writing the reports and creating more accurate diagnoses.

C. Method The majority of the methods used to study the task of image captioning for medical imaging such as: [3, 4, 5, 6, 7, 8] are based on deep learning models that use an image-encoder text-decoder architecture. The encoder, a Convolutional Neural Network (CNN), is used to encode the features of the images and the decoder (a Recurrent Neural Network (RNN) such as LSTM [9] or, more recently, a transformer [10]) is then used to convert the extracted features into text. A pure transformer based encoder-decoder in the context of medical image captioning tasks has not, however, been studied in depth, until recent in [11].

GIT [12] is an encoder-decoder based transformer, that has been pre-trained on 0.8B (billion) image-text pairs from numerous sources such as the datasets: COCO [13], Conceptual Captions(CC3M) [14], Conceptual Captions (CC12M) [15], SBU [16], Visual Genome [17], ALT200M [18] where it obtained state-of-the-art results for various tasks including image captioning. However, GIT has not been previously tested on any medical imaging datasets and tasks. Our proposed solution refers to using and modifying the GIT transformer for the challenging task of automated report generation for radiography images.

D. Results The results obtained after testing the proposed solution have been comparable with the state-of-the-art models that use either transformer based architectures or Convolutional Neural Network/Recurrent Neural Network architectures, on natural language generation metrics (RougeL [19], Meteor[20]), as well as classification metrics (F1 macro and F1 micro).

2 Related Work

A. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation [21] A common problem that emerges for the task of image captioning for medical images is the generation of incomplete, inconsistent and thus clinically inaccurate reports. Tackling this issue is detrimental to generating correct, relevant and conclusive medical reports that will automatically improve the quality of the generated reports and provide help for clinicians to state faster and more accurate diagnoses.

The majority of approaches that tackle the task of image captioning for medical images base their methods on encoder-decoder architectures. Similarly, Miura et al. [21] adopts a comparable system that can be observed in figure 2. The images associated with the report are passed through an image encoder in order to extract the meaningful features of the x-ray that will be given to a text decoder that will generate the report, creating the final prediction. The architecture behind this method consists mainly on the use of the M^2 Trans, originally proposed in [22], a meshed-memory transformer for image captioning that in this case is extended to multiple images, as seen in figure 1.

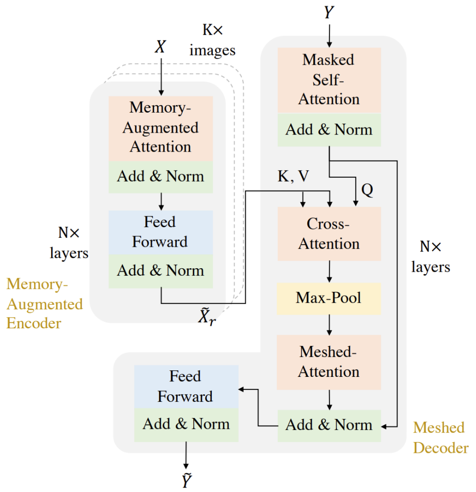
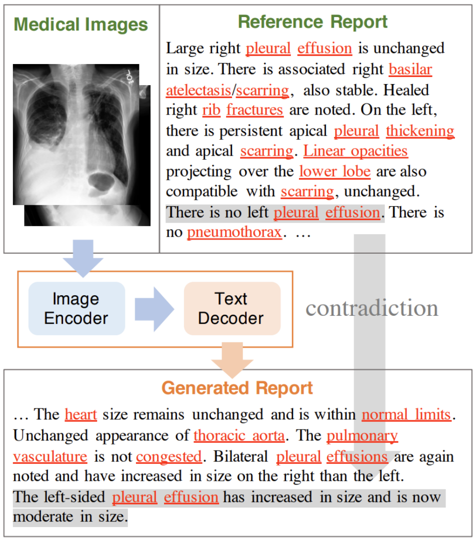


Figure 1: The M^2 Trans architecture extended to multiple images.



Large right pleural effusion is unchanged in size. There is associated right basilar atelectasis/scarring, also stable. Healed right rib fractures are noted. On the left, there is persistent apical pleural thickening and apical scarring. Linear opacities projecting over the lower lobe are also compatible with scarring, unchanged. There is no left pleural effusion. There is no pneumothorax. …

… The heart size remains unchanged and is within normal limits. Unchanged appearance of thoracic aorta. The pulmonary vasculature is not congested. Bilateral pleural effusions are again noted and have increased in size on the right than the left. The left-sided pleural effusion has increased in size and is now moderate in size.

Figure 2: Example of workflow of the described system in order to generate a report.

The work of Miura et al. [21] especially differentiates itself through the addition of two rewards that encourage the factual completeness and consistency of radiology report generation. The two rewards introduced are Exact Entity Match Reward (factENT) and Entailing Entity Match Reward (factENTNLI). The metric factENT measures the amount of radiology domain components present in a generated report, compared to a target report, in order to cover as much medical information from the report as possible, thus assessing its completeness. The metric factENTNLI is an extension of the factENT with a Natural Language Inference (NLI) model in order to consider the differences in meaning between the generated reports at inference and the reference reports.

The method is tested on two different datasets, one of which is MIMIC-CXR [23], obtaining better results than previous state-of-the-art papers such as [24]. The metrics used in order to assess the efficacy and accuracy of the method are both natural language generation (NLG) metrics such as Bleu4, CIDEr-D and BERTscore, and classification metrics such as the F1 micro score calculated on the most frequent 5 out of the 14 radiological label categories of the CheXBert Labeler [25] - *atelectasis*, *cardiomegaly*, *consolidation edema*, *pleural effusion* [23]. However, as NLG metrics such as CIDEr or Bleu4 may create incomplete and inconsistent generations, the F1 score is used in order to determine the clinical accuracy of the generated reports. The F1 score is computed by using the CheXBert Labeler [25], a radiology report labelling method based on a biomedically pretrained BERT [26] that has near radiologist performance in labeling medical conditions. CheXBert has the purpose of extracting and classifying medical reports into 14 common diseases.

B. Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model [27] CXR-RePaiR is a retrieval-based chest radiography report generation method that has obtained good results on the natural language generation metric BLEU [28], s_{emb} semantic similarity metric and F1 macro score, surpassing previous state-of-the-arts such as [24] and [21]. The approach described in this paper can be seen in the figure below 3.

CXR-RePaiR is based on an encoder-decoder architecture. From the medical report corpus, a pre-trained text encoder extracts the key sentences of the report. From an input x-ray image, the image encoder (a Convolutional Neural Network - CNN) conceals the image features creating a latent representation that will then be associated with the previously generated text encodings. A prediction is made by selecting an image representation and text pair that has the best similarity. The created prediction is passed through the CheXBert Labeler [25] that creates a one-hot classification of the prediction. Similarly, for the target reports, the CheXBert Labeler is also used for the classification. In the end, the two classifications (from the prediction and the target) are used to compute a classification performance score, F1 macro, for all the 14 radiological label categories of the CheXBert Labeler [25]. The CheXBert Labeler is a radiology report labeling method based on a biomedically pretrained BERT [26] that extracts and classifies medical reports into 14 common diseases. CheXBert has near radiologist performance in labeling medical conditions.

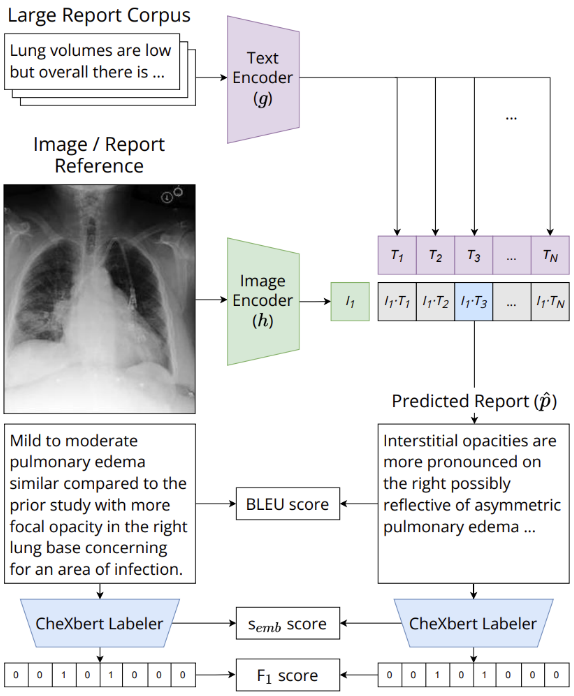
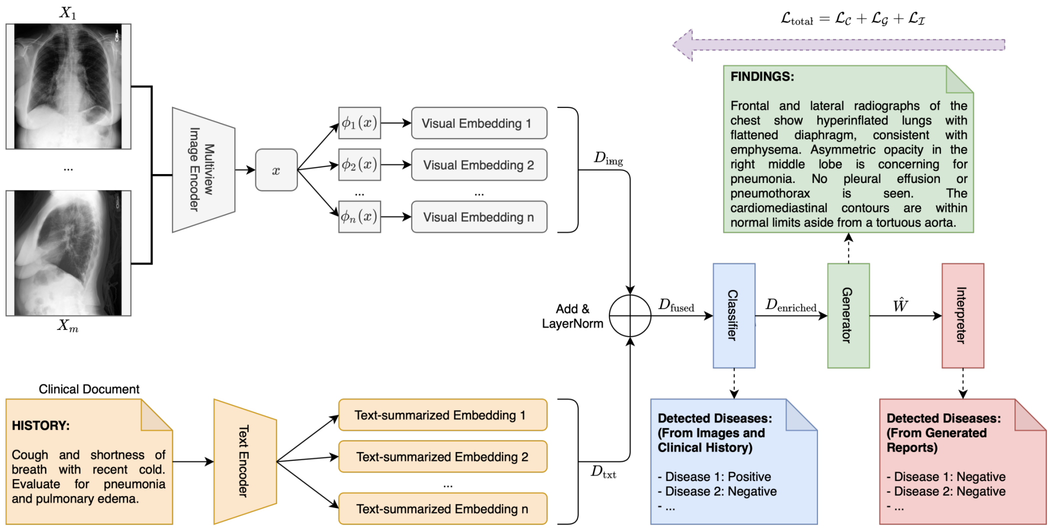

Large Report Corpus
Lung volumes are low but overall there is ...
Text Encoder (g)
...
Image / Report Reference
Image Encoder (h)
I_1
T_1
T_2
T_3
...
T_N
$I_1{\cdot}T_1$
$I_1{\cdot}T_2$
$I_1{\cdot}T_3$
...
$I_1{\cdot}T_N$
Predicted Report ($\hat{p}$)
Mild to moderate pulmonary edema similar compared to the prior study with more focal opacity in the right lung base concerning for an area of infection.
BLEU score
Interstitial opacities are more pronounced on the right possibly reflective of asymmetric pulmonary edema ...
CheXbert Labeler
s_{emb} score
CheXbert Labeler
0 0 1 0 1 0 0 0
F_1 score
0 0 1 0 1 0 0 0

Figure 3: The CXR-RePaiR architecture.

An important mention is that s_{emb} metric is computed between the prediction embeddings and the target embeddings obtained with the CheXBert Labeler. This score represents the cosine similarity (semantic similarity) of the predicted report in relation to the target report. This forces the model during training to accept only the predictions that have a very big similarity with the original reports by using the same model (CheXBert Labeler) that later will be used to compute the F1 score. This is done in order to increase the F1 score computed at inference. Apart from this, the NLG metric Bleu2 is also used between the target report and prediction report.

C. Automated Generation of Accurate and Fluent Medical X-ray Reports [29] The work of Nguyen et al. [29] proposes a different approach from what we have previously researched, basing its method on a structure of three modules: the classification module, the generator module and the interpreter module. The goal of this approach is to produce clinically accurate and fluent reports, addressing a well known issue of inconsistency and incompleteness of medical report generation. The architecture of the system can be seen in the figure 4.

The x-ray images are first passed through a multi-view image encoder that creates visual embeddings of the features of the images. Synonymous, the clinical documents (reports) are passed through a text encoder that creates text-summarised embeddings of the reports. The image and text embeddings are then passed through three complementary modules. The classification module learns the disease feature representation, accounting the disease related topics and creating enriched disease embeddings. The embeddings created by the classifier module are then given to a transformerbased generator that reconstructs them into medical reports. Finally, the interpreter module reads the reports generated and fine-tunes them, increasing the consistency between the classification of the generated reports and the classification given by the first module.



FINDINGS:

Frontal and lateral radiographs of the chest show hyperinflated lungs with flattened diaphragm, consistent with emphysema. Asymmetric opacity in the right middle lobe is concerning for pneumonia. No pleural effusion or pneumothorax is seen. The cardiomediastinal contours are within normal limits aside from a tortuous aorta.

HISTORY:

Cough and shortness of breath with recent cold. Evaluate for pneumonia and pulmonary edema.

Detected Diseases: (From Images and Clinical History)

- Disease 1: Positive
- Disease 2: Negative
- ...

Detected Diseases: (From Generated Reports)

- Disease 1: Negative
- Disease 2: Negative
- ...

Figure 4: The architecture of the described method.

This system achieves impressive results on both NLG metrics such as BLEU [28], RougeL[19] and Meteor[20] on two different datasets, one of which being MIMIC-CXR [23].

D. Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach [11] Compared to previous works that use CNN-based architectures for image features extraction, the work of Wang et al. [11] is the first proposed method using a pure transformer-based framework for medical x-ray report generation. It has obtained impressive results over natural language generation metrics such as BLEU [28] and RougeL [19], over two chest radiography datasets, including MIMIC-CXR [23]. Its architecture can be seen in figure 5. The system aims at solving two very common challenges that come with this task: 1. because chest x-ray images are very much alike, the important features of the image reside in small details and 2. critical medical information/words usually is surrounded by commonly used descriptions in a clinical report which makes the model brush over these details and misconstrue them for unimportant entities.

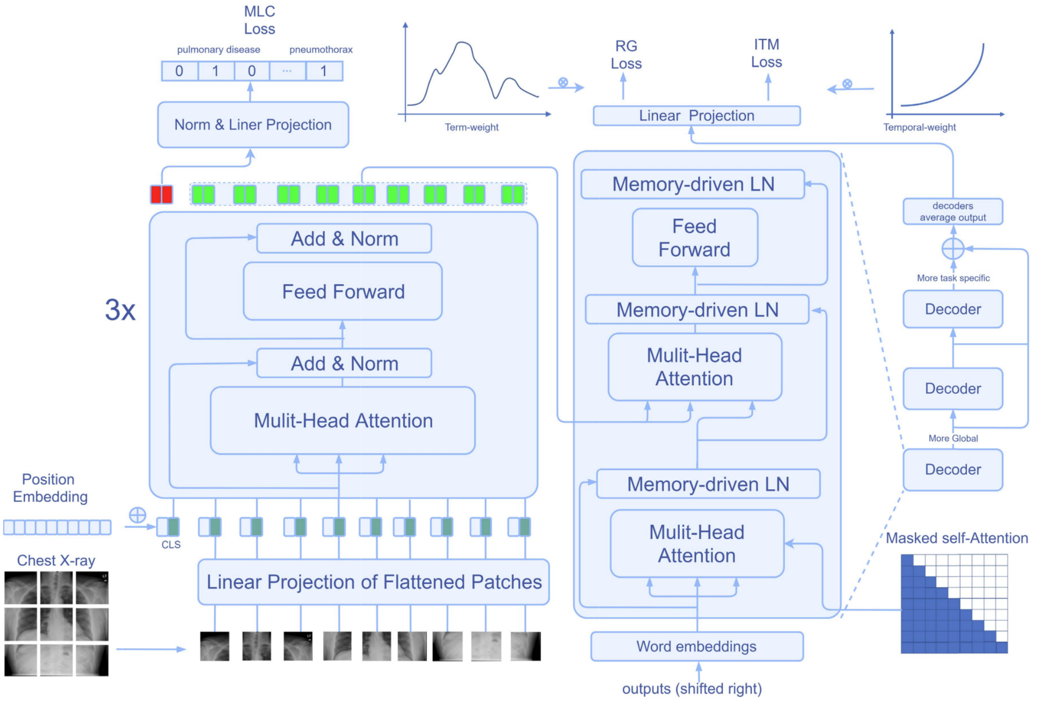


Figure 5: The architecture of the described method.

The proposed solution for these problems is a pure transformer-based framework that is capable of multi-label diagnostic classification and weighting the importance of the entities of the report. This architecture, as seen in the figure below 5, is composed of an encoder-decoder transformer structure that has an image-text matching mechanism added in order to tackle the first problem of image similarity by helping the model better correlate image and text features. Additionally, a multilabel classification task is integrated into the framework in order to guide the model into making accurate predictions and solve the second problem stated above. Moreover, a separate scheme is constructed to have the ability of distinctively weighing and representing the key words during training thus training the model to identify important information that can be easily overlooked.

3 Proposed Method

We propose a transformer based encoder-decoder approach by using the GIT Transformer (GIT: A Generative Image-to-text Transformer for Vision and Language [12]) on the MIMIC-CXR dataset [23], in order to generate the medical reports.

GIT is a generative image-to-text transformer that has obtained state-of-the-art results on various computer vision tasks such as image captioning, question answering, video captioning and decent results on image classification and scene text recognition. It has been pre-trained on 0.8B (billion) image-text pairs from various sources such as the datasets: COCO [13], Conceptual Captions(CC3M) [14], Conceptual Captions (CC12M) [15], SBU [16], Visual Genome [17], ALT200M [18], but has not been previously tested on any medical imaging datasets and tasks. Our proposed solution refers to using and modifying the GIT transformer for the challenging task of automated report generation for radiography images.

The architecture of the GIT transformer can be seen in the figure 6. An image encoder (based on the model of [30]) extracts the features of the input image and creates a compact 2D feature map, flattened into a list of features which are fed to the text decoder (a transformer module) that uses both the encodings and the previously tokenized and embedded text description of the image in order to predict the text description.

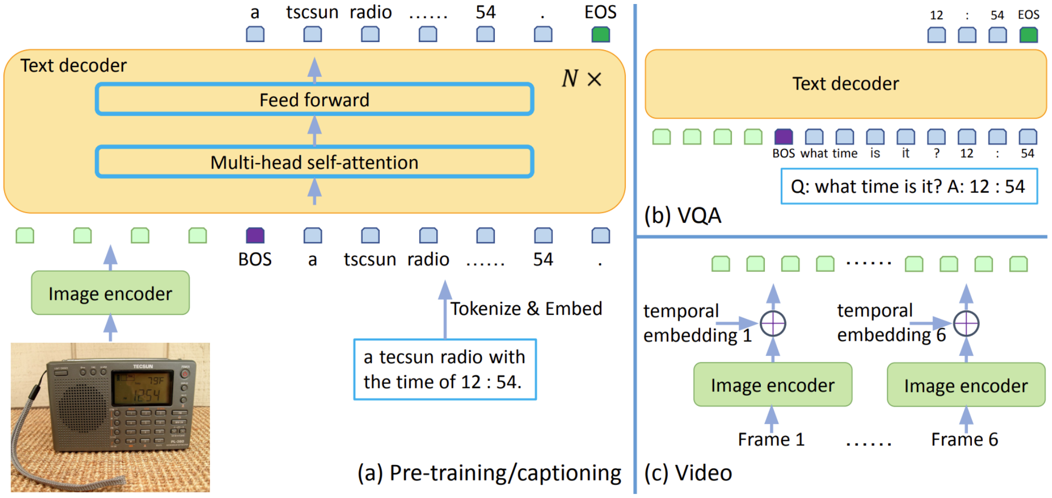


Figure 6: The GIT architecture. On the left part(a), the image captioning architecture is described.

Apart from just using the GIT Transformer, we also experiment with introducing an auxiliary loss for multi-label classification. Similar to [11], the output of the vision encoder is passed through linear projection and normalization layers. Then, we employ a classification head for each of the 14 possible diagnostics and define the classification loss as the mean of the losses computed for each head via weighted Cross Entropy, as shown in Equation 1, where D is the number of diagnostics (in our case 14), x is the input image, $f_v(x)$ is the image encoding, given by passing x_i through the vision encoder f_v, $h_i(f_v(x))$ is the prediction of the i-th classification head, y_i is the target for the diagnostic i and CE is the Cross Entropy loss:

$$\mathcal{L}_{MLC} = \frac{1}{D} \sum_{i=1}^{i \leq D} CE(y_i, h_i(f_v(x))) \tag{1}$$

4 Experiments

A. Environment The experiments were conducted on a NVIDIA GEFORCE RTX2080 GPU. One experiment takes around two and a half days to complete for 15 epochs. The MIMIC-CXR [23] dataset required one month in order to fully download on an external hard disk drive of 7TB due to its size and large number of necessary HTTP requests.

B. Model GIT has been extended to many variants out of which we have chosen to test GIT base (GITb [12]). GITb is a smaller version of GIT, that has been pre-trained on a total of 10M image-text pairs, or 4M images of size 224x224 extracted from the datasets COCO [13], Conceptual Captions(CC3M) [14], SBU [16] and Visual Genome [17]. The model size of GIT base is 129M parameters, much smaller when compared to GIT that has 681M parameters. The main reason we decided to conduct our experiments using the GITb variant is the resource limitations of the available GPU's as the other variants are much larger and would need more time to run the experiments and more VRAM memory. Additionally, even though the larger models have shown better results on the image captioning task, the differences in the results shown between the four variants (GIT base, GIT large, GIT and GIT2 [12]) are not sufficiently large to constitute a valid reason to use a more costly option than GIT base. The GITb version used by us is the one fine-tuned on the COCO dataset.

C. Dataset The MIMIC-CXR dataset (MIMIC Chest X-Ray Database v2.0.0 [23]) is the largest publicly available dataset containing chest x-ray images with their corresponding, free-text, clinical reports. The dataset contains a total of 377,110 DICOM format radiography images that correspond to 227,835 studies of 64,588 patients. These studies have been performed at the Beth Israel Deaconess Medical Center in Boston, Massachusetts, examined between 2011 - 2016.

The MIMIC-CXR dataset containes 10 folders (p10-p19). One folder containes approximately 6,500 patients. One patient can have one or more studies and one study can have more than one image. One study can contain comparison, clinical history, indication, reasons for examination, impressions (where a general outlook on the report is given in one sentence), and findings (a detailed diagnosis of the patient's condition is made). Different works use different sections of the studies, for example Endo et al. [27] uses both the findings and the impressions. We decided to use only the findings section of the MIMIC-CXR dataset following the work of Nguyen et al. [29] and Miura et al. [21].

The possible radiography images are chest x-ray images taken from the front (anterior-posterior (AP)), back (posterior-anterior (PA)), lateral (LA), or more particularly, left-lateral (LL) part of the patient. Their distribution can be seen in figure 7

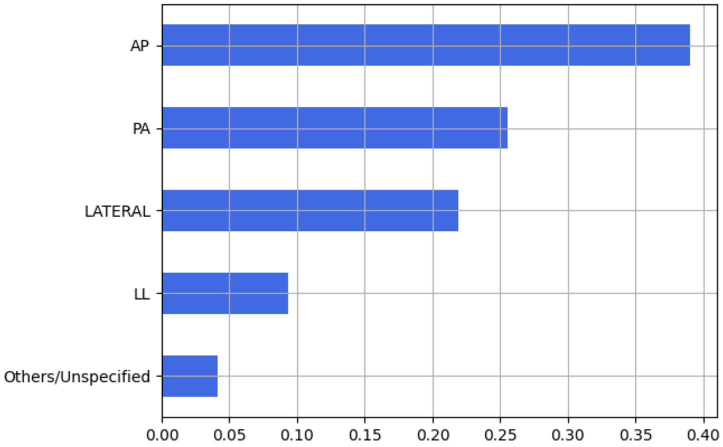


Figure 7: The distribution of the possible view positions of the x-ray images.

The reason why we are keeping just the AP and the PA images is because we do not use multiple images at once (like [29] and[21]), correlated to a report. We keep one image at a time (similar to [27] and [11]) that corresponds to a report and we chose to keep the AP and PA because these images are the ones that contain the most information about the condition of a patient. If the report would state that a tumor is seen in the left lung and we would have kept a single lateral image taken from the side, the model would not be able to correctly identify the tumour.

For our experiments, we have used only 7 folders out of 10. After dropping the studies that do not contain the findings section, and dropping the images that are not AP or PA, the total amount of images that we have used is 116,594.

The splits were made in a 80% training or 93,210 images, 11,692 images for the validation set (10%) and the rest 10% of images for the test set, another 11,692. However, the evaluation of the model was taking a very long time because the NLG metrics are computed on a generated report which required the model to be applied in an autoregressive fashion. For this reason, we ended up reducing the validation set to only 200 images. Because the test set is used only once, at the end, and not for every epoch, we were able to keep the test set to its original size. After all of these optimizations, the training time decreased to approximately 2 days for one experiment trained for 15 epochs.

D. Evaluation Metrics The metrics used to evaluate the proposed method are both NLG metrics such as BLEU, RougeL and Meteor as well as the classification metrics F1 macro and F1 micro in order to determine the factual completeness and clinical accuracy as well.

Prior works in the domain of clinical report generation use two types of metrics for: first, natural language generation (NLG) metrics are used, as this is the default evaluation approach used for image captioning tasks in order to evaluate the models' ability to generate coherent text. Second, as Boag et al. [31] found that it is possible to have models with high NLG scores that don't produce correct diagnosis, clinical accuracy metrics have been introduced in order to measure the ability of the model to produce reports that would lead to the right diagnosis [21, 24, 31].

For the NLG metrics, different works use different metrics. For example, Endo et al. [27] uses Bleu2, Miura et al. [21] reports Bleu4, while Wang et al. [11] and Nguyen et al. [29] use all the Bleu scores up to 4 n-grams. Additionally, the two of them also compute the RougeL metric, while Meteor is only reported by [29].

Similarly, for the clinical accuracy metrics, while multiple works agreed on using the CheXbert labeler [25] for obtaining predictions and computing the F1 score, this was implemented in various ways. For example, Miura et al. [21] passes the generated reports through the CheXbert labeler in order to obtain labels for the 5 most frequent diagnostics and then computes the F1 micro for these predictions against the ground truth labels. On the other hand, Endo et al. [27] uses the CheXbert labeler to obtain labels for the generated reports on all 14 radiological categories and then computes the F1 macro between the predictions and the ground truth labels.

In order to be able to compare with all these previous approaches, we employ the union of the metrics used by them in order to evaluate our approach. This results in using Bleu1, Bleu2, Bleu3, Bleu4, RougeL and Meteor as natural language generation metrics and using both F1 macro (on 14 diagnostics) and F1 micro (on 5 diagnostics) as clinical accuracy metrics.

BLEU [28] is a NLG metric that computes the n-gram precision (i.e. the fraction of n-grams from the evaluated text that are found in the target text) for each order (e.g. for Bleu4, the precision is computed for 1-grams, 2-grams, 3-grams and 4-grams) and computes the geometric average of the computed precisions. Then, BLEU factorizes a so called Brevity Penalty, which is a factor meant to penalize the evaluated text for being shorter than the targets. This aims to compensate for not taking recall into account, as it can't be clearly defined when a set of multiple potential targets is available.

ROUGE [19] is a NLG metric that computes n-gram recall (i.e. the fraction of n-grams from the reference text that are found in the evaluated text). While the recall can't be clearly defined when using a set of potential targets, what ROUGE does is to compute a score for each potential target, and then average all the scores. However, this introduces the drawback of not being able to obtain a perfect score when having multiple distinct targets. This and other disadvantages are discussed in detail by Schluter [32]. There are multiple types of ROUGE available, for example Rouge-n which uses n-grams for computing the recall, or RougeL, which uses the longest common subsequence instead. We use RougeL for evaluating our model, similar to [11] and [29].

METEOR [20] is a metric introduced for machine translation which aims to overcome the limitations of other NLG metrics such as BLEU or ROUGE. The main idea is to create a unigram matchings (a.k.a. alignment) between the evaluated text and the reference text, such that each unigram is matched to at most one unigram in the other text. Each match doesn't necessarily have to be an exact match, as matchers based on the stemmed forms or on the meanings can also be used. For an alignement, the METEOR is computed by combining the precision, recall and fragmentation (i.e. how well the unigrams in the evaluated text are ordered, in relation to the reference). In the case of multiple targets, the maximum score is being considered, thus solving the problem described above for the ROUGE metric.

As METEOR provides a way of combining both precision and recall and it is also better correlated with human judgement [20], we choose METEOR to be the metric we use for selecting the best model on the validation set, before computing the performance on the test set.

5 Results and Discussion

A. Analysis The results on the MIMIC-CXR dataset can be seen in the table 1. We compare our method with various state-of-the-art systems previously explained in the Related Work section of this document. GIT-CXR is the unmodified version of the GIT approach where as GIT-CXR-CLS is the approach with the classification task introduced.

We can observe that the best results were obtained with the original approach, GIT-CXR, with the results of GIT-CXR-CLS being at around 0.05 smaller that the unmodified method. This difference is not big, but it is surprising as following the previous approaches that use the same idea of an added classification task such as [11], have reported improvements in their results.

One potential reason for this behaviour could be the fact that we kept 4 labels for each classification head (i.e. each diagnostic), corresponding to "positively mentioned", "negatively mentioned",

“mentioned with uncertainty” and “no mention”. On the other side, previous work using this approach [11] only kept the positive and negative labels. Retrospectively, this would make more sense, as introducing a separate class for "uncertain" or for "unknown" might confuse the model and not allowing the vision encoder to learn useful representations for the image. Additionally, we didn’t experience with different weighting schemes between the two losses. From the loss graphs we could observe how the MLC loss is about 4x higher than the main loss, which might make the model get stuck into trying to optimize the auxiliary loss, thus overlooking the generation loss. This assumption is also supported by the fact that the generated reports tend to be shorter when introducing the MLC loss.

MODEL	BL1	BL2	BL3	BL4	RL	M	F1MACRO	F1MICRO
GIT-CXR (OURS)	0.254	0.131	0.081	0.055	0.256	**0.271**	**0.474**	0.552
GIT-CXR-CLS (OURS)	0.157	0.084	0.055	0.038	0.238	0.242	0.372	0.507
ARR [11]	0.351	0.223	0.157	0.118	0.287	–	–	–
AGA [29]	**0.495**	**0.360**	**0.278**	**0.224**	**0.390**	0.222	–	–
CXR-REPAIR [27]	–	0.217	–	–	–	–	0.274	–
M^2 TRANS [21]	–	–	–	0.133	–	–	–	**0.567**

Table 1: Results on the MIMIC-CXR dataset [23]. The best results for each task are highlighted using **bold** font. Compared to the other methods showed in this table that have used 100% of the MIMIC-CXR dataset, our experiments have been completed on only 70% of the dataset. The F1 macro is determined on all 14 of the radiological label categories of the CheXBert Labeler [25], following the work of [27]. The F1 micro is determined on only 5 of the 14 diseases of CheXBert, following the work of [21].

Apart from the differences in results between the GIT-CXR-CLS and the GIT-CXR, we can also observe some discrepancies between the performance of our model (GIT-CXR) and the performance of prior works.

Firstly, on the NLG metrics, we can see that all the BLEU metrics are considerably lower for our model. This is in contrast to the other NLG metrics, as RougeL is comparable to ARR[11] (although smaller than AGA[29]), while METEOR is better than that of AGA[29], although AGA obtains the best performance on all other NLG metrics.

Secondly, on the clinical accuracy metrics, we can observe that our model obtains an F1 micro score comparable to the other reported performance of [21] , while the F1 macro score of our model is substantially higher than the other reported result of [27]. Unfortunately, clinical accuracy metrics are not available for the other models.

Next, we analyze why our model performs so well on the clinical accuracy metrics and a few NLG metrics, while it has a very poor performance on other metrics, especially for the BLEU scores.

By looking into the generated reports, we can observe that they tend to be significantly shorter than the targets. More precisely, while a target report has a mean length of 56.66 tokens, a generated report is expected to have on average only 36.82 tokens, which account for only 64.3% of the reference text. The full distributions for the length of the generated reports, the length of the target reports and their comparison can be observed in Figures 8, 9 and 10, respectively.

As the generated reports are too short, the brevity penalty used by the BLEU score is low (0.57), which has a high impact on all the BLEU scores. In the case of GIT-CXR-CLS, the generated reports are even shorter (46.4% of the target), which results in a lower brevity penalty of about 0.31, severely impacting the BLEU scores. Additionally, a shorter generated report can only have a negative impact on the recall-based metrics (as there will be less n-grams to match the ones in the target), so it is only natural for the RougeL and METEOR to also be affected.

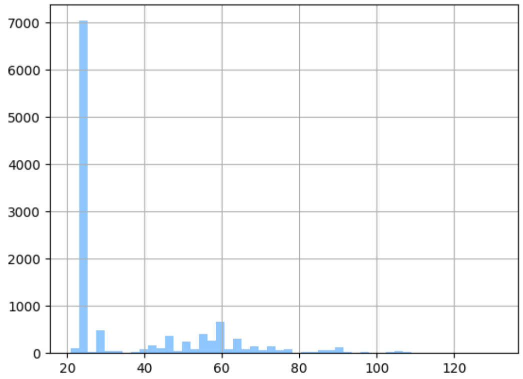


Figure 8: Distribution of generated reports length.

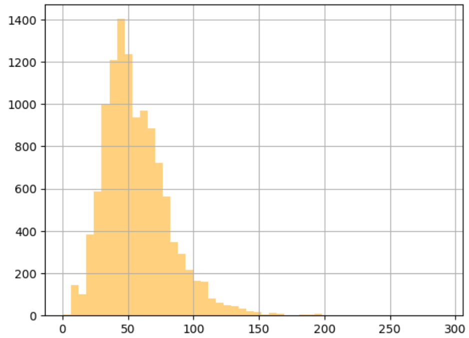


Figure 9: Distribution of target reports length.

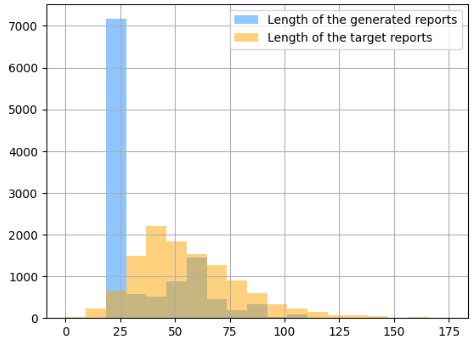


Figure 10: Both target and generated reports length in comparison

Another thing to analyze is the diversity of the generated reports. As it can be seen in Figure 8, there is a concentration of generated reports with the length of 24, which might imply a lack of diversity. Indeed, by analyzing the generated reports we reach the conclusion that 51.5% of them actually consist of a single repeating instance, which is:

- *the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.*

By analyzing the test targets for which this report is generated, we also get reports that would have the diagnostic of "no findings", some of them repeating multiple times, the second of them actually being the exact match of the generated report:

- *Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.* - 69 times
- *The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.* - 36 times
- *The lungs are clear. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities.* - 25 times
- etc.

This behaviour may actually explain the discrepancy between high clinical accuracy scores (as the model correctly identifies a large amount of the radiography images without findings only by using this report) and the lower NLG scores (as it is not ideal to use the same report from a generation perspective).

However, by looking into the other reports, we can observe how this behaviour of low diversity generated reports does not apply to the rest of the test set, as the frequency of the generated reports drops from 51.5% for the top result, to 8.2% for the second, 3.4% for the third, followed by a long tail, which actually results in a mean frequency of only 0.003% of the test set, despite these top results that make up a big portion of the dataset.

B. Resource Limitations Another thing to consider when analyzing the results is the limited number of setups we were able to experiment up until now, due to physical constraints such as available GPUs, the running time of an experiment the time and space needed for downloading the dataset, etc.

This not only affected our ability to test multiple setups (e.g. a weighting scheme between losses, in the case of introducing the MLC loss; testing additional changes to the architecture or to the training procedure), but may also impacted the results of the current setups.

For example, in Figure 14 we can see that the ratio between the length of the generated reports and the target reports was still increasing, so allowing the model to train for a longer time might result in more qualitative reports (e.g. longer reports would decrease the impact of the brevity penalty while computing the BLEU scores; longer reports would also favor diversity). Figure 15 also shows that the model was still able to learn. This is also supported by Figures 11, 12 and 13, as all the metrics had a global ascending tendency even on the validation dataset, despite the steep local changes in direction. The graphs could be smoother and more informative if we were to use more samples for validation, but this also requires more training time, as the inference needed for validation is very slow.

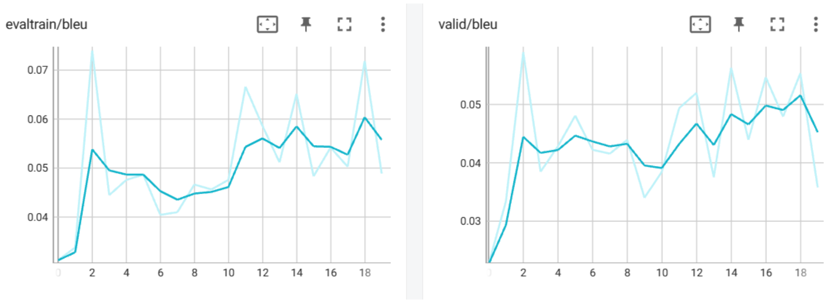


Figure 11: BLEU scores during training (left) and validation (right).

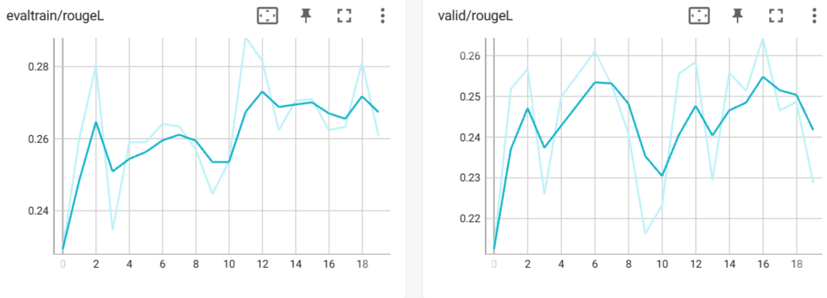


Figure 12: RougeL scores during training (left) and validation (right).

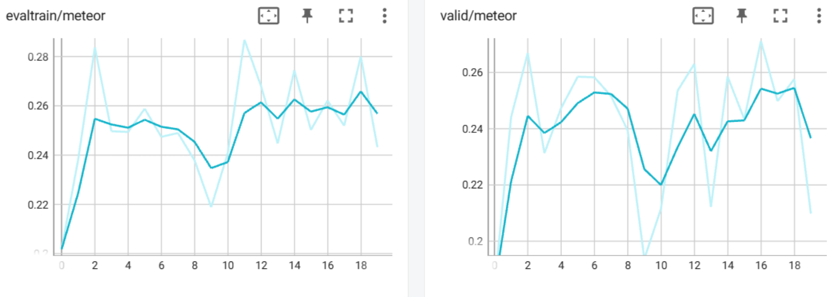


Figure 13: Meteor scores during training (left) and validation (right).

While the time constraints didn't allow for extensive experimentation, all these limitations could definitely be addressed in future work, as we already shown the feasibility of the approach and now it is mostly a matter of time and computing resources.

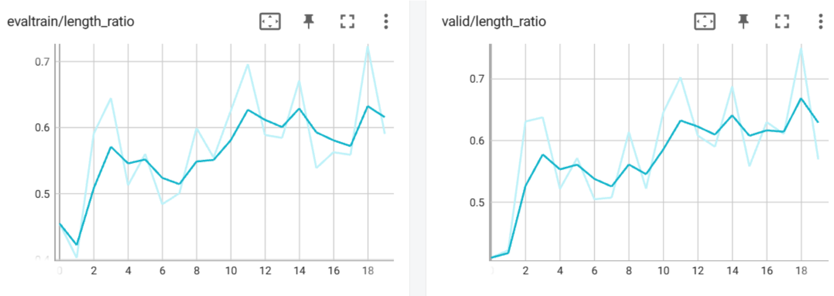


Figure 14: Length ratio between the generated reports and the targets during training (left) and validation (right).

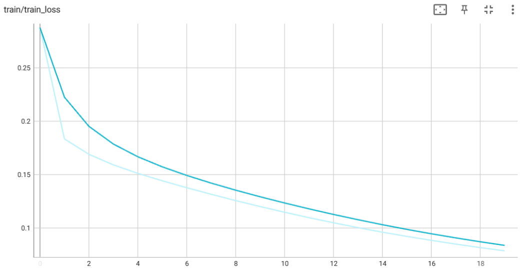


Figure 15: Train loss.

6 Conclusion

Although many previous approaches based on encoder-decoder systems, composed of a CNN image-encoder and RNN text-decoder, or more recently also transformer based methods have been previously proposed with good results, the task of automated medical report generation for chest radiography images is far from being solved as it poses many complex problems to this day. Radiography images are composed of important detailed features that can be easily missed and require fine-grained approaches of encoding, key words residing into the general information of reports can also be overlooked, and the generation of clinical reports is usually factually incomplete or clinically inaccurate for various reasons. In this work, we have studies these existing problems as well as tested our own method of dealing with such challenges with the use of the GIT Transformer. The results obtained, even with just 70% of the dataset have been comparable to current state-of-the-art methods and have proven that pure transformer-based architectures are worth researching forward.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[3] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[4] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018.

[5] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673, 2019.

[6] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE international conference on data mining (ICDM)*, pages 728–737. IEEE, 2019.

[7] Justin Lovelace and Bobak Mortazavi. Learning to generate clinically coherent chest x-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, 2020.

[8] Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, 41(10):2803–2813, 2022.

[12] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[15] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[16] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[18] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022.

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[20] Satanjeev Banerjee and Alon Lavie. 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, 2005.

[21] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.

[22] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

[23] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[24] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

[25] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[29] Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*, 2021.

[30] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[31] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *Machine learning for health workshop*, pages 126–140. PMLR, 2020.

[32] Natalie Schluter. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–45. Association for Computational Linguistics, 2017.