Generating X-ray Reports Using Global Attention

Felipe André Zeiser[1], **Cristiano André da Costa** [1], **Gabriel de Oliveira Ramos** [1], **Henrique C. Bohn** [1], **Ismael Santos** [1], **Bruna Donida** [2], **Ana Paula de Oliveira Brun**[2], **Nathália Zarichta**[2]

[1]Software Innovation Laboratory - SOFTWARELAB
Applied Computing Graduate Program
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Brazil

[2]Gerência de Ensino e Pesquisa
Grupo Hospitalar Conceição
Porto Alegre, Brazil

Abstract. *The use of images for the diagnosis, treatment, and decision-making in health is frequent. A large part of the radiologist's work is the interpretation and production of potentially diagnostic reports. However, they are professionals with high workloads doing tasks operator dependent, that is being subject to errors in case of non-ideal conditions. With the COVID-19 pandemic, healthcare systems were overwhelmed, extending to the X-ray analysis process. In this way, the automatic generation of reports can help to reduce the workload of radiologists and define the diagnosis and treatment of patients with suspected COVID-19. In this article, we propose to generate suggestions for chest radiography reports evaluating two architectures based on: (i) Long short-term memory (LSTM), and (ii) LSTM with global attention. The extraction of the most representative features from the X-ray images is performed by an encoder based on a pre-trained DenseNet121 network for the ChestX-ray14 dataset. Experimental results in a private set of 6650 images and reports indicate that the LSTM model with global attention yields the best result, with a BLEU-1 of 0.693, BLEU-2 of 0.496, BLEU-3 of 0.400, and BLEU-4 of 0.345. The quantitative and qualitative results demonstrate that our method can effectively suggest high-quality radiological findings and demonstrate the possibility of using our methodology as a tool to assist radiologists in chest X-ray analysis.*

1. Introduction

The Coronavirus pandemic challenged the world in the most diverse aspects [Huang et al. 2020]. Each country has built its way of combating SARS-CoV-2, from implementing public policies to choosing ways of diagnosis and treatments during the pandemic [Zeiser et al. 2022]. The world, in this context, created and shared health-related information in a way and speed never seen before to save the largest number of people infected by the virus in each country. These data are, today, a source for the creation of new approaches to the disease treatment and diagnosis producing new tools for healthcare [Wong et al. 2020].

The gold standard for diagnosing COVID-19 is the Quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR) test [Patel et al. 2020,

Huang et al. 2020]. However, other exams, such as chest X-ray, have high sensitivity for findings of the disease, such as intraparenchymal consolidations, and ground glass [Wong et al. 2020]. In the context of public health in Brazil, the delay in performing and reporting RT-qPCR affects early diagnosis as a way of infection control and patient care, suggesting other ways as an alternative to detect the disease and estimate prognosis [Zeiser et al. 2022]. Furthermore, the chest X-ray, a simple and low-cost exam, proved to be an essential exam in the screening and fight against coronavirus, being able to perform the diagnosis of pneumonia and record the evolution of the disease through serial images [Wong et al. 2020]. Therefore, with the opportunity to use X-rays as a way for diagnosis and prognosis for the disease rapidly spreading around the world.

From this perspective, artificial intelligence has proved to be a highly valuable tool in processing large volumes of data and demonstrating the potential to assist health professionals in medical decisions [Lakhani and Sundaram 2017, De Fauw et al. 2018]. Furthermore, developing computational architectures based on Convolutional Neural Network (CNN) models capable of making a report suggestion with high accuracy and high positive predictive value in a short time can help decision-making in the health context.

Therefore, this article explores the application of Deep Learning (DL) techniques in generating suggestions of radiological reports in X-ray images of patients with COVID-19. The method consists of a pre-trained encoder/decoder architecture on a chest X-ray set. In addition, we introduced a new dataset composed of X-rays and reports collected from a public hospital. The experimental results suggest that the proposed methodology is a helpful tool to aid in the process of chest X-ray analysis. The main contributions are as follows:

- We propose a neural network structure for generating chest X-Ray report suggestions. The model generates a suggestion for an X-Ray report by extracting features from the images using a convolutional architecture. The characteristics are then processed along with the tokens already provided by an LSTM network with attention.
- Our experiments demonstrate consistent performance improvements when adding our proposed modules. In addition, our model achieves state-of-the-art performance for the dataset used.

The rest of this paper is organized as follows. In Section 2, we present the most significant related works for the definition of the work. Section 3 presents the methodology of the work. Section 4 details the results. Finally, Section 5 presents the conclusions of the work.

2. Related Work

We evaluated suspected regions in the images (X-rays) for radiological diagnoses. Radiologists are trained to assess various features and identify radiological findings according to X-ray features and clinical history. These findings are then reported in the radiological report. In this way, the radiological report summarizes the clinical findings and represents a rich information for the patient's staging [Granata et al. 2021]. In this way, training machine learning models to suggest findings in natural language format can provide interpretability to the models. Recently, methods based on extracting representations with convolutional neural networks for generating natural language using Long

short-term memory (LSTM) or attention mechanisms have attracted some research interest [Zhang et al. 2017].

Over the years, several methods have been proposed for the visual captioning task, producing just one sentence describing the contents of the images [Xu et al. 2015, You et al. 2016, Rennie et al. 2017]. However, these methods were limited to practical application in generating radiological reports. In this sense, more sophisticated proposals were proposed with better capabilities to provide details of findings in X-ray images. One of the forms currently used is the division of the report into two groups: (i) findings and (ii) impressions. Findings consist of a detailed description of the information identified in the radiograph. Meanwhile, the impression section forms the conclusion of the diagnosis based on the information presented in the findings section [Jing et al. 2020].

Another alternative is the mining of tags that can represent most of the findings presented in chest X-ray [Shin et al. 2016]. However, transforming the natural language generation into a classification process can reduce the explainability and comfort of radiologists in interpreting the result [Chen et al. 2020]. Therefore, methods that can generate the report suggestion with characteristics similar to those used by radiologists present a higher degree of feasibility for practical application [Jing et al. 2020]. The proposal of integrated methods, which can generate natural language and detect findings in the images, despite the current low performance, presents an exciting approach to assist in understanding changes in X-ray [Wang et al. 2018].

A hierarchical model can be one of the ways to overcome challenges in the X-ray chest report generation process. At first, the main findings' detection is performed, combined with the visual resources, and fed back into an LSTM network to generate the report descriptions [Jing et al. 2017]. Using transformers using models with memory-augmented mechanisms helps to improve the ability to generate long coherent text sequences. However, the proposed model showed a dominant behavior in generating normal findings [Chen et al. 2020].

In summary, several recent works have investigated the automatic generation of reports for X-ray chest images. However, the labels of current methods are mainly limited to diseases and do not present contextual information. In addition, the datasets mainly present normal cases, so it is challenging to detect disease cases and rare findings in unbalanced data. Therefore, these works lack evidence of their ability to generalize the problem, making their use as an aid system for radiologists unfeasible. Thus, this work has as its main contribution the evaluation of two methods for generating report suggestions. In addition, we introduced a new dataset with a balanced distribution of cases collected from a public hospital.

3. Materials and Methods

An overview of the methodology employed in this work is presented in Fig. 1. The methodology is divided into four stages: preprocessing, data augmentation, training, and testing. Preprocessing consists of image resizing, contrast normalization, and report processing (Section 3.2). The data augmentation step describes the methods used to generate synthetic images (Section 3.3). In the training stage, the models and the parameters used are defined (Section 3.4).

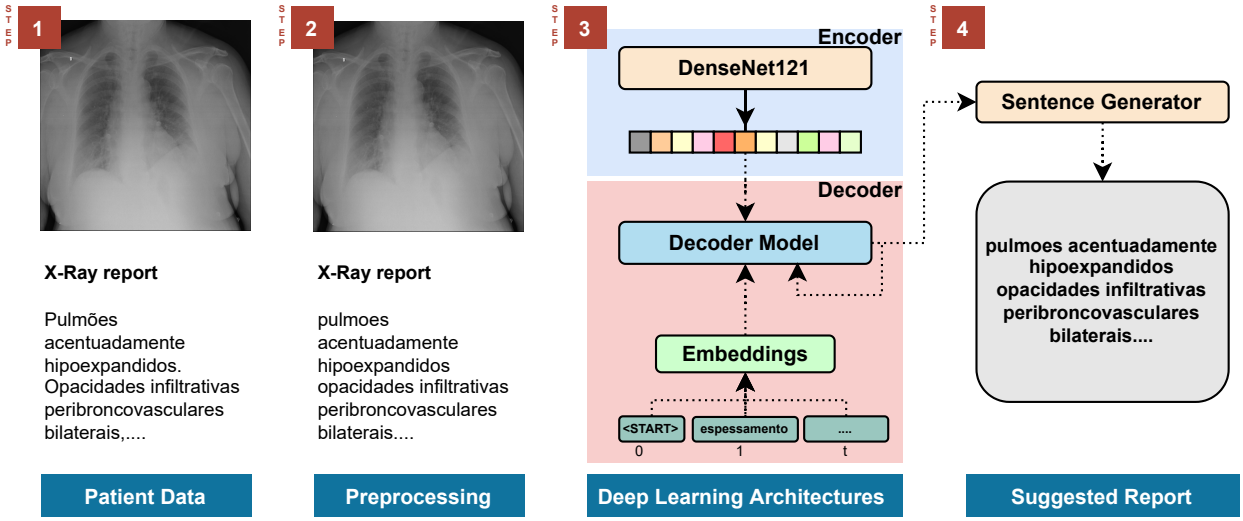


Figure 1. Diagram of the proposed methodology. The first step consists of collecting the X-rays and the radiological reports. In the second step, the X-ray images and the reports go through a pre-processing step, normalizing the images and removing irrelevant tokens for inference. Finally, step three is responsible for extracting X-ray characteristics for generating the report suggestion presented in step four.

3.1. Dataset

We collected 6650 chest radiographs from patients with COVID-19 and without COVID-19 from a hospital in Porto Alegre. The information collected comprises clinical data and radiographic reports from 697 patients. In this work, we used 3280 images of patients with COVID-19 and 3370 images without COVID-19.

3.2. Pre-processing

In this step, due to the available computational capacity, the radiographs were resized to 512x512 pixels. The reduction was proportional in each axis, with a black border added to the axis with the smallest size. Finally, we applied Contrast-Limited Adaptive Histogram Equalization (CLAHE), which proved to be efficient in improving the performance of DL [Pooch et al. 2020] algorithms. CLAHE subdivides the image into sub-areas using interpolation between the edges. To avoid noise increase, uses a threshold level of gray, redistributing the pixels above that threshold in the image. CLAHE can be defined by:

$$p = [p_{max} - p_{min}] * G(f) + p_{min} \quad (1)$$

where p is the pixel's new gray level value, the values p_{max} and p_{min} are the pixels with the lowest and highest values low in the neighborhood and $G(f)$ corresponds to the cumulative distribution function [Zuiderveld 1994]. In Fig. 2, we present an example with the original and preprocessed image.

In addition, we kept only alpha characters, removing special characters, numbers, and punctuation from reports before the tokenization process to reduce the number of tokens. As a result, the maximum size of tokens in the reports was 40 tokens and defined

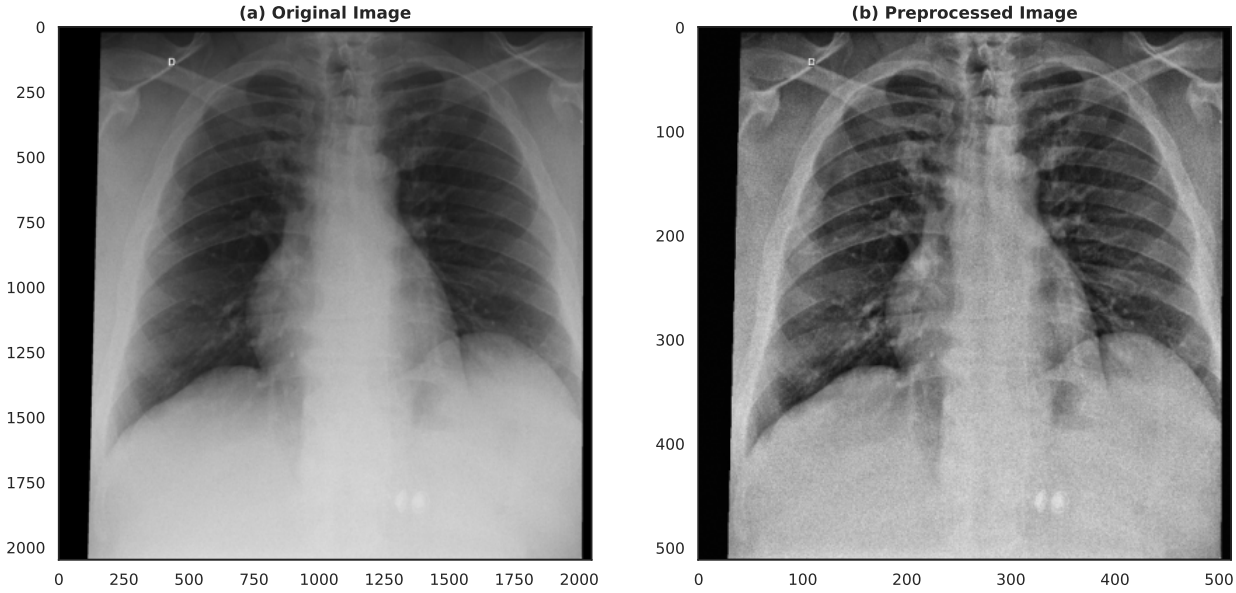


Figure 2. (a) Original XR image; (b) Preprocessed image.

considering the 80 percentile value. Finally, the radiographs and reports were divided randomly into three sets, training (70%), validation (10%), and testing (20%).

3.3. Data Augmentation

For a better generalization in the training process, we performed a balancing of our data set. In this way, we counted the number of radiographs with the same reports and performed a horizontal flip of those with less than two occurrences. The option of performing only the horizontal flip is motivated by the characteristics of a chest X-ray, which presents slight variations that may indicate the presence or absence of specific pathologies. This way, transformations such as distortions or shearing can add noise to the image, such as the deformation of organs, making the report inference process difficult.

3.4. Deep Learning Architectures

For the generation of the report suggestions, we used a methodology based on image captioning composed of an image encoder and a decoder. We evaluated two architectures, both of which employed as encoder a DenseNet121 pre-trained on the ChestX-ray14 dataset [Wang et al. 2017]. ChestX-ray14 is one of the largest open datasets, containing over 100,000 chest X-ray images. The decoders used were based on: (i) LSTM; and (ii) Global attention with LSTM. For the LSTM, we used a set of 512 LSTM units followed by a dense layer of 1567 units, corresponding to our dictionary size. Finally, in the model with Global attention, we used 512 LSTM units followed by a global attention layer and a dense layer of 1567 units. In Fig. 3 we present a view of the neural network architectures used to generate the report suggestion.

In the training stage, we used the sparse categorical loss as the loss function, considering only the losses for the words present in the original report. To optimize the weights of the models, we use Adam's algorithm. At the end of each epoch, we use the validation set to measure the model's accuracy on an independent set and obtain the best training weights.

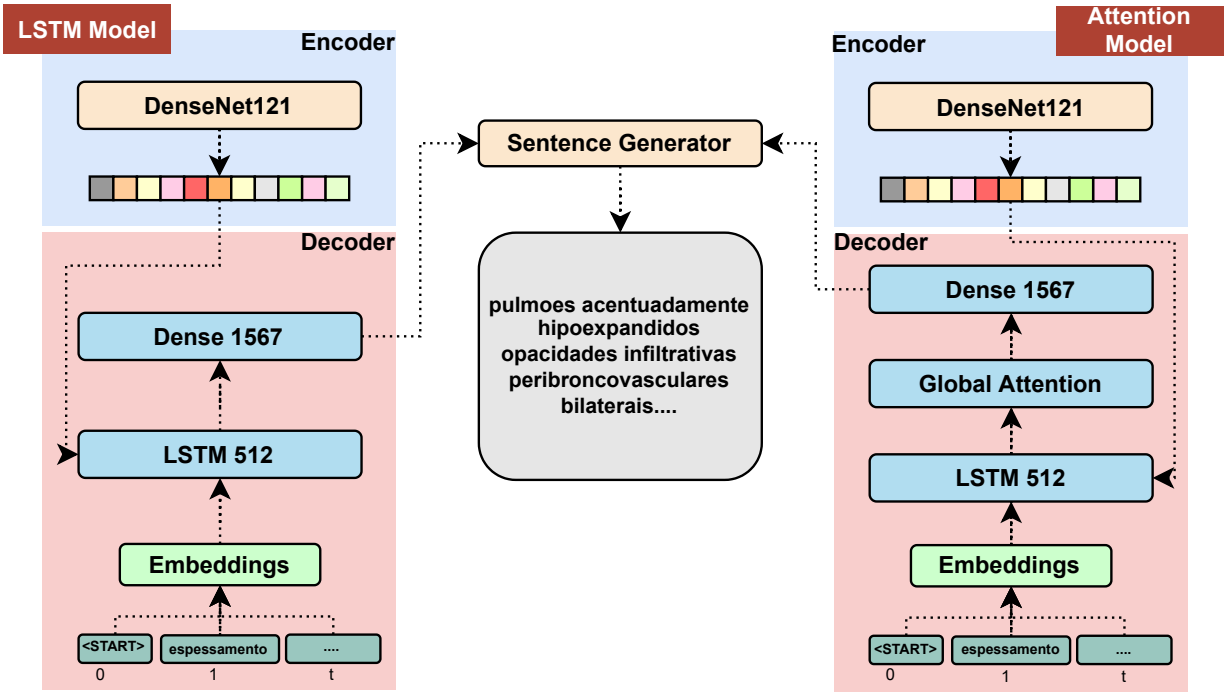


Figure 3. Architecture of neural networks for the image features extraction and report suggestion generation..

3.5. Evaluation Metrics

The performance evaluation of the report suggestion architecture will be performed using the Bilingual Evaluation Understudy (BLEU) metric. BLEU compares each predicted word in the sentence with the reference sentence. In other words, BLEU metrics are determined by comparing a candidate sentence with reference sentences in n-grams [Aafaq et al. 2019]. The BLEU score can be calculated by:

$$BLEU = min(1 - \frac{lr}{lc}, 0) + \sum_{n=1}^{N} w_n \log p_n \quad (2)$$

where l_r and l_c is the ratio between the size of the reference and predicted sentence, w_n are positive weights, and p_n is the geometric mean of the modified n-gram accuracies [Aafaq et al. 2019].

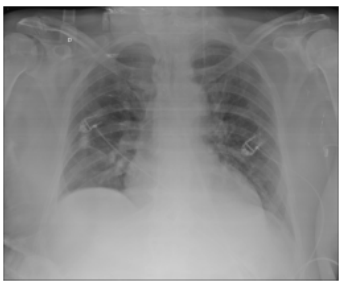
4. Experiment Results and Discussion

In this section, we present and evaluate the proposed models' results. We train the models for 100 epochs for each set. We evaluate each model at the end of training in the validation set. The choice of the best set of weights was performed automatically based on the error for the validation set. Tab. 1 presents the values obtained for the evaluation metrics in the test set.

Analyzing the results of the models, we can see stability between the BLEU performance metrics. However, the LSTM model presented inferior results compared to the model with Global Attention. This result may be related to the functioning of the LSTM

Table 1. Results for the test set for each model.

Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
LSTM	0.470	0.304	0.219	0.165
Global Attention	0.693	0.496	0.400	0.345



True Report: pulmoes pouco insuflados nao se observa sinal de lesao consolidativa ou tumescente cateter vascular esquerda com extremidade distal projetada na topografia da veia cava superior tubo endotraqueal area cardiaca aumentada aorta com calcificacoes parietais seios costofrenicos laterais livres
Predicted Report: nao ha evidencia de lesao tumescente ou consolidacao no parenquima pulmonar area cardiaca aumentada aorta com calcificacoes parietais seios costofrenicos laterais livres presenca de cateter esquerda

Figure 4. Example of a chest X-ray processed by LSTM model with Global Attention.

mechanism. The LSTM network will take the previous output as the contextual representation of all the information already processed. In this way, the LSTM networks have difficulty preserving the dependency for longer than some steps, making it challenging to maintain the context in the process report inference [Cho et al. 2014]. Meanwhile, the attention mechanism manages to preserve the context and measure weight for each output already processed, which can facilitate the process of identifying pathologies in X-ray images.

To illustrate the performance of the proposed methodology, we present the results of the inference process of the reports in Fig. 4. In Fig. 4, we present a case of a lung that, according to the radiologist's report, does not present adequate inflation, which may indicate a patient's respiratory difficulty. In addition, it is possible to observe a change in the cardiac area and calcifications in the aorta, which may indicate cardiovascular disease. When we compare the report generated by the model, we can see an ability to detect small changes in the images. In addition, we can highlight two findings, the presence of a catheter in the left X-ray area and the identification of calcifications in the patient's aorta. Finally, given the variability of the images, the models still have some problems differentiating some radiological findings, such as the exact position of the catheter or the fact that the lungs are not adequately inflated.

Finally, we compare our results using the attention mechanism with LSTM and the encoder based on DenseNet121. It is essential to highlight that our work is based on a dataset with reports written in Portuguese, which makes it difficult to compare directly with the works proposed in the current literature. However, the results obtained for the proposed method are comparable to the state-of-the-art for the suggestion of chest X-ray reports. In Tab. 2, we present the main results obtained by the related works.

Study	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4
[Shin et al. 2016]	OpenI	**0.793**	0.091	0.0	0.0
[Jing et al. 2017]	IU X-Ray	0.517	0.386	0.306	0.217
[Wang et al. 2018]	OpenI	0.239	0.124	0.086	0.065
[Chen et al. 2020]	IU X-RAY	0.470	0.304	0.219	0.165
[Jing et al. 2020]	CX-CHR	0.693	**0.626**	**0.580**	**0.545**
Our method	Private	0.693	0.496	0.400	0.345

Table 2. Comparison of our method with the related works.

5. Conclusion

This article compared two pre-trained models for the ChestX-ray14 dataset for the automatic suggestion of chest X-ray reports. To improve the generalizability of the results, we applied a set of pre-processing techniques. The main scientific contribution of this study was the proposal of an architecture for the generation of chest radiological reports. The pre-trained models can serve as a basis for future studies and provide a second opinion to the radiologist during X-ray analysis.

The present study has some limitations. First, the current results are based on a single institution dataset and transfer learning, achieving performance comparable to gold standard expert assessment. Therefore, the current method still requires improvements in detecting less frequent findings and learning based on a set of reports that considers a broader range of writing modes of different radiologists. Furthermore, this can provide more insights into the daily chest X-ray analysis protocol.

As future work, we intend to expand our dataset with a collection of chest X-rays and reports from other institutions participating in an established research consortium. In addition, we hope to analyze the influence of public datasets on the image features learning process of chest X-ray with COVID-19, analyzing whether the proposed models can generalize features to different datasets. Furthermore, to provide a more significant explanation of the method, we intend to add an attention mechanism to the image inference process indicating which areas were more activated. Another point that can be explored is the use of Transformers mechanisms for the decoder and encoder of the model. In addition, further studies are needed regarding the use of the Portuguese language in generating report suggestions. This is because the Portuguese language and the process of writing the report differ considerably from the standard adopted in the English language. Finally, a proposal for a multimodal methodology, for example, using clinical data, laboratory tests, and images, can be helpful in the transparency of diagnostic recommendations.

References

Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., and Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350.

Granata, V., Pradella, S., Cozzi, D., Fusco, R., Faggioni, L., Coppola, F., Grassi, R., Maggialetti, N., Buccicardi, D., Lacasella, G. V., et al. (2021). Computed tomography structured reporting in the staging of lymphoma: A delphi consensus proposal. *Journal of clinical medicine*, 10(17):4007.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506.

Jing, B., Wang, Z., and Xing, E. (2020). Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.

Jing, B., Xie, P., and Xing, E. (2017). On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

Lakhani, P. and Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582.

Patel, A., Jernigan, D. B., et al. (2020). Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak—united states, december 31, 2019–february 4, 2020. *Morbidity and mortality weekly report*, 69(5):140.

Pooch, E. H. P., Alva, T. A. P., and Becker, C. D. L. (2020). A deep learning approach for pulmonary lesion identification in chest radiographs. In *Brazilian Conference on Intelligent Systems*, pages 197–211. Springer.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. (2016). Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised clas-

sification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058.

Wong, H. Y. F. et al. (2020). Frequency and distribution of chest radiographic findings in patients positive for covid-19. *Radiology*, 296(2):E72–E78.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

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.

Zeiser, F. A., Donida, B., da Costa, C. A., de Oliveira Ramos, G., Scherer, J. N., Barcellos, N. T., Alegretti, A. P., Ikeda, M. L. R., Müller, A. P. W. C., Bohn, H. C., et al. (2022). First and second covid-19 waves in brazil: A cross-sectional study of patients' characteristics related to hospitalization and in-hospital mortality. *The Lancet Regional Health-Americas*, 6:100107.

Zhang, Z., Xie, Y., Xing, F., McGough, M., and Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436.

Zuiderveld, K. (1994). Graphics gems iv. In Heckbert, P. S., editor, *Graphics Gems*, chapter Contrast Limited Adaptive Histogram Equalization, pages 474–485. Academic Press Professional, Inc., San Diego, CA, USA.