Two-stage Conditional Chest X-ray Radiology Report Generation

Pablo Messina [1,5,6], José Cañete [2,6], Denis Parra [1,5,6],
Álvaro Soto [1,6], Cecilia Besa [3,5], and Jocelyn Dunstan [4,5]
[1] Department of Computer Science, Pontifical Catholic University of Chile.
[2] Department of Computer Science, University of Chile.
[3] Department of Radiology, School of Medicine, Pontifical Catholic University of Chile.
[4] Center for Mathematical Modeling (CMM), University of Chile.
[5] Millennium Institute for Intelligent Healthcare Engineering (iHEALTH), Chile.
[6] National Center for Artificial Intelligence (CENIA), Chile.
{pamessina,cbesa}@uc.cl, {dparra,asoto}@ing.puc.cl,
jose.canete@ug.uchile.cl, jdunstan@uchile.cl

Abstract

A radiology report typically comprises multiple sentences covering different aspects of an imaging examination. With some preprocessing effort, these sentences can be regrouped according to a predefined set of topics, allowing us to implement a straightforward two-stage model for chest X-ray radiology report generation. Firstly, a topic classifier detects relevant findings or abnormalities in an image. Secondly, a conditional report generator outputs sentences from an image conditioned on a given topic. We present experimental results on the test split of the MIMIC-CXR dataset for each stage separately and the system as a whole. Most notably, the proposed model outperforms previous works on several medical correctness metrics based on the CheXpert labeler, establishing a new state-of-the-art. The source code is available at https://github.com/PabloMessina/MedVQA/.

1 Introduction

The development of deep neural networks for image-based radiology report generation is a very active research area given the large number of medical examinations and a continuing shortage of trained radiologists worldwide [19, 13]. In particular, the limitations of conventional NLP metrics (e.g. BLEU[23], ROUGE-L [16]) have inspired several publications to explore more adequate metrics to evaluate the clinical accuracy of machine-generated reports [20, 4, 17, 31, 25, 24, 19]. In this context, the CheXpert labeler [11] has emerged as a standard tool for evaluating medical correctness in chest X-ray report generation. We contribute to this area with a novel yet simple model for two-stage conditional report generation. The first stage identifies relevant topics and the second stage generates sentences for the report conditioned on a given topic. We evaluate our model on the test split of the MIMIC-CXR dataset [12], outperforming previous works on several metrics based on the CheXpert labeler.

2 Methods and Results

Defining a set of topics. Through an iterative incremental approach via regular expressions and random sampling, we examine the sections *findings* and *impression* of informative reports from MIMIC-CXR [12] and IU X-ray [8] and mine a vocabulary of 97 topics (e.g. diseases such as *fibrosis* and *COPD*, abnormalities such as *adenopathy*, and general topics such as *lungs*, *heart*, *bones* and

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

tubes-and-lines). Sentences in a report that match a regular expression associated with a given topic become part of the ground truth for that topic. This means that the same sentence could match more than one topic, and multiple topics can be present in a report. Additionally, inspired by the template-based approach by Pino et al. [24], we use the CheXpert labels as 14 additional topics, with two possible templates for each one (see C in Figure 1). Likewise, we extend this idea to other datasets, namely, CheXpert [11] (14 labels), CXR14 [30] (14 labels) and VinDr-CXR [21] (28 labels), with two templates for each label.

Topic Classifier (TC). The first step in order to build a report is to select topics. There are many options here, e.g. select all topics, a predefined subset, etc. Another option is to implement a Topic Classifier neural network for the 97 mined topics (A in Figure 1). A fully connected layer predicts mined topics based on 3 sources of information: (1) visual encoder's global features (we use DenseNet121 [10]), (2) the patient's clinical history or indication (available in most reports in MIMIC-CXR and IU X-ray), encoded by a BiLSTM and decoded back by an LSTM to improve the encoding quality (autoencoder loss), and (3) a weighted sum of 14 CheXpert embedding vectors (to exploit any correlations), where the weights come from the sigmoid activations of a multilabel classification layer that predicts CheXpert labels from the image as an auxiliary task.

Conditional Report Generator (CRG). B in Figure 1 shows CRG's architecture. DenseNet121 performed the best as visual encoder. However, we also experimented with a Vision Transformer from CLIP [26], which was later fine-tuned on MIMIC-CXR using contrastive loss in order to test the effect of visual-language multimodal pre-training. We supervise the global features of the visual encoder with multiple auxiliary tasks, depending on the metadata available in each dataset. Topics are represented with a topic embedding matrix. Global and local features from the visual encoder are fed along with the topic vector to a Transformer Decoder [29] that generates a mini-report word by word. Given multiple topics, the final report is simply the concatenation of the respective mini-reports.

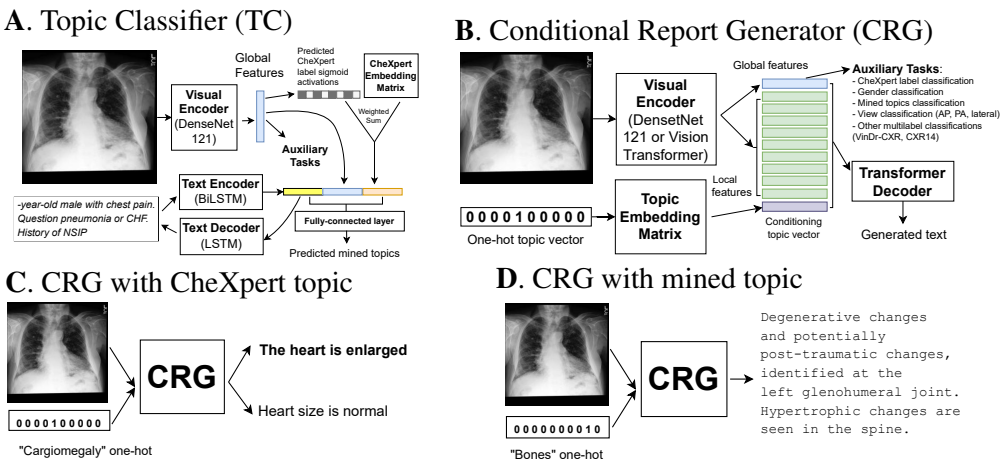

Figure 1: Architectures of the Topic Classifier (TC) and Conditional Report Generator (CRG) networks. **A** shows the TC architecture. **B** shows the CRG architecture. **C** shows CRG conditioned on a CheXpert topic, where the output can be either a pre-defined positive or negative sentence (as in Pino et al. [24]). **D** shows CRG conditioned on a mined topic (the output is more verbose).

2.1 Results

New state-of-the-art. Table 1 presents results of CRG with 3 different topic selection schemes: *chexpert*, *predicted by TC ensemble* and *ground truth*. All three schemes consistently achieve higher CheXpert macro recall and F1 scores than previous works, although only the first two are fully automatic. The *ground truth* scheme works as an oracle that obtains the topics directly from the ground-truth reports, so these results represent a theoretical upper bound with an optimal TC module. Considering the fully automatic results only, our best macro F1 score was $0,477$ (row 15, with CheXpert topics) and our best micro F1 score was $0,588$ (row 17, predicted mined topics). From our literature review, to the best of our knowledge our closest competitors are Pino et al. [24] (row 7) with a macro F1 score of $0,428$, and Nguyen et al. [22], with a micro F1 score of $0,576$. Interestingly,

both works explicitly include a multilabel classification task to predict CheXpert labels, as we also do in this work, suggesting that this is a key decision to achieve higher CheXpert labeler metrics in report generation.

Table 1: Report generation results on the test split of MIMIC-CXR. Bold indicates the highest score in a group. Red indicates the highest score overall. Notation used: DN = DenseNet 121; TF = Transformer; ViT = Vision Transformer; M = MIMIC-CXR; I = IU X-ray; Ch = CheXpert; C14 = CXR14; V_{test} = VinDR-CXR (test split); V_{all} = VinDR-CXR (all data); ft = fine-tuned; fve = frozen visual encoder; medtok = medical tokenization (only medical terms, stopwords are ignored); B = BLEU; R-L = ROUGE-L; C-D = CIDEr-D; F1 = F1 score; P = precision; R = recall.

ID	Model	NLP			Med.	CheXpert (Macro)			CheXpert (Micro)		
		B	R-L	C-D	Comp.	F1	P	R	F1	P	R
	Other works										
1	Liu et al. [17]	0.192	0.306	**1.046**	-	0.180	0.313	0.126	0.334	**0.634**	0.227
2	Chen et al. 2020 [7]	0.205	0.277	-	-	0.276	0.333	0.273	-	-	-
3	Chen et al. 2021 [6]	0.208	0.283	-	-	0.303	0.352	0.298	-	-	-
4	Lovelace et al. [18]	0.257	0.318	0.316	-	0.228	0.333	0.217	0.441	0.475	0.361
5	Miura et al. [20]	-	-	0.509	-	0.304	0.361	0.360	0.563	0.499	**0.646**
6	Nguyen et al. [22]	**0.339**	**0.390**	-	-	0.412	**0.432**	0.418	**0.576**	0.567	0.585
7	Pino et al. [24]	0.094	0.185	0.238	-	**0.428**	0.381	**0.531**	-	-	-
8	Kong et al. [14]	0.243	0.286	-	-	-	-	-	0.519	0.482	0.563
	Our work										
9	$CRG(DN+TF)_{\text{chexpert topics}}$: M	0.146	0.196	**0.041**	0.087	0.464	0.377	0.713	0.557	0.428	0.797
10	$CRG(DN+TF)_{\text{chexpert topics}}$: M+I	0.146	0.196	**0.041**	0.087	0.469	0.388	0.678	0.569	0.448	0.781
11	$CRG(DN+TF)_{\text{chexpert topics}}$: M+I+Ch	0.146	0.196	**0.041**	0.088	0.463	0.384	0.689	0.568	0.446	0.783
12	$CRG(DN+TF)_{\text{chexpert topics}}$: M+I+Ch+C14	0.146	0.196	0.040	0.088	0.463	0.386	0.702	0.564	0.440	0.785
13	$CRG(DN+TF)_{\text{chexpert topics}}$: M+I+Ch+C14+$V_{test}$	0.145	0.195	0.040	0.088	0.467	0.386	**0.712**	0.569	0.439	**0.811**
14	$CRG(DN+TF)_{\text{chexpert topics}}$: M+I+Ch+C14+$V_{all}$	0.145	0.195	**0.041**	0.088	0.462	0.383	0.700	0.571	0.444	0.800
15	$CRG(DN+TF)^{\text{medtok, fve, ft}}_{\text{chexpert topics}}$: M+I	0.146	0.197	0.040	0.086	**0.477**	0.392	0.693	0.575	0.449	0.799
16	$CRG(ViT_{CLIP}+TF)^{\text{medtok, fve, ft}}_{\text{chexpert topics}}$: M+I	**0.150**	**0.199**	0.040	0.087	0.472	0.389	0.653	0.582	0.464	0.779
17	$CRG(DN+TF)^{\text{medtok, vmf, ft}}_{\text{mined topics predicted by ensemble}}$: M+I	0.102	0.184	0.031	**0.116**	0.448	**0.400**	0.568	**0.588**	**0.487**	0.743
18	$CRG(DN+TF)_{\text{gt mined topics}}$: M	**0.185**	**0.299**	0.031	0.166	0.587	0.559	0.649	0.692	0.639	0.756
19	$CRG(DN+TF)_{\text{gt mined topics}}$: M+I	0.180	0.295	0.035	0.162	0.572	0.540	0.629	0.685	0.631	0.748
20	$CRG(DN+TF)_{\text{gt mined topics}}$: M+I+Ch	0.176	0.292	0.023	0.159	0.586	0.548	0.657	0.690	0.621	0.777
21	$CRG(DN+TF)_{\text{gt mined topics}}$: M+I+Ch+C14	0.180	0.296	0.033	0.162	0.592	0.551	0.660	0.689	0.629	0.762
22	$CRG(DN+TF)_{\text{gt mined topics}}$: M+I+Ch+C14+$V_{test}$	0.180	0.293	0.025	0.157	0.594	0.557	0.654	0.698	0.635	0.776
23	$CRG(DN+TF)_{\text{gt mined topics}}$: M+I+Ch+C14+$V_{all}$	0.180	0.293	0.029	0.160	0.583	0.550	0.650	0.685	0.627	0.755
24	$CRG(DN+TF)^{\text{medtok}}_{\text{gt mined topics}}$: M+I+Ch+$V_{all}$	0.160	0.260	0.152	0.171	0.592	0.567	0.639	0.704	0.653	0.762
25	$CRG(DN+TF)^{\text{medtok}}_{\text{gt mined topics}}$: M+I+Ch+C14+$V_{all}$	0.167	0.269	**0.179**	**0.183**	0.621	**0.604**	0.666	**0.729**	**0.670**	**0.798**
26	$CRG(DN+TF)^{\text{medtok, fve, ft}}_{\text{gt mined topics}}$: M+I	0.165	0.264	0.153	0.177	**0.623**	0.596	**0.676**	0.719	0.656	0.797
27	$CRG(ViT_{CLIP}+TF)^{\text{medtok, ft}}_{\text{gt mined topics}}$: M+I+Ch+C14+$V_{all}$	0.167	0.268	0.170	0.181	0.598	0.569	0.650	0.703	0.649	0.768

CheXpert topics vs. mined topics. CheXpert-based reports achieve superior CheXpert macro scores (except for precision), whereas the two-stage model with a TC ensemble achieves superior CheXpert micro scores (except for recall). Additionally, we implement a simple "medical completeness" metric that measures n-gram overlaps between medical terms (ignoring non-medical terms), revealing that reports based on mined topics cover more medical terms than CheXpert-based reports that are limited to 28 templates.

Contrastive pre-training. Rows 16 and 27 in Table 1 and row 8 in Table 2 present our best results with a dual encoder of Vision Transformer (pre-trained from CLIP [26]) and a Bio Clinical BERT [2] fine-tuned on MIMIC-CXR using contrastive loss. Despite this pre-training, we failed to observe significant improvements compared to variations based on DenseNet121 without this pre-training. This could be due to the smaller scale of medical datasets, where CLIP's modality-specific attention might not perform as well as other multimodal techniques [9].

Visual encoder results. An interesting result in Table 2 is that the CheXpert classification metrics of visual encoders trained end-to-end with the rest of the CRG architecture are consistently superior to those of TC models. We hypothesize that this could be due to CRG's visual encoder benefiting from the additional natural language generation supervision. We also observe a high agreement between the outputs of the visual encoder and the Transformer conditioned on CheXpert topics, measured by Cohen's Kappa. This suggests that the Transformer decoder is learning to take visual encoder features into account rather than merely exploiting topic biases (a problem in other works observed by Babar et al. [3]). On the other hand, TC models achieved superior performance in mined topic classification, and an ensemble of four TC models performed the best.

Impact of additional datasets. We test adding datasets during training with a multi-dataset data loader to see any generalization improvements on MIMIC-CXR. We use IU X-ray [8], CheXpert [11], CXR14 [30], and VinDr-CXR [21]. We run multiple experiments, including curriculums, where a model is trained on multiple datasets and later fine-tuned on a subset. Overall, we only managed to

Table 2: Visual encoder results on the test split of MIMIC-CXR. For CRG models, CheXpert metrics for the Transformer decoder when conditioned on CheXpert topics are included. Cohen's Kappa measures the agreement between visual encoder and Transformer.

ID	Model	Mined Topics		CheXpert (visual encoder)				CheXpert (transformer)		
		F1 (macro)	F1 (micro)	ROC-AUC (macro)	ROC-AUC (micro)	F1 (macro)	F1 (micro)	F1 (macro)	F1 (micro)	Cohen's Kappa
1	CRG(DN+TF): M	0.213	0.413	0.758	0.823	0.473	0.578	0.465	0.556	0.703
2	CRG(DN+TF): M+I	0.208	0.409	0.750	0.821	0.465	0.579	**0.476**	0.570	0.767
3	CRG(DN+TF): M+I+Ch	0.206	0.392	0.763	0.821	0.472	0.576	0.474	0.569	0.785
4	CRG(DN+TF): M+I+Ch+C14	0.209	0.401	**0.765**	0.823	0.478	0.582	0.470	0.563	0.806
5	CRG(DN+TF): M+I+Ch+C14+V_{test}	0.212	0.405	0.761	0.826	0.483	0.587	**0.476**	0.570	0.785
6	CRG(DN+TF): M+I+Ch+C14+V_{all}	0.210	0.405	0.762	**0.829**	0.481	0.587	0.475	0.572	0.806
7	CRG(DN+TF)^{medtok}: M+I+Ch+C14+V_{test}	0.216	**0.422**	**0.765**	**0.829**	**0.486**	**0.593**	0.473	0.569	0.781
8	CRG(ViT$_{CLIP}$+TF)$^{medtok,\ fve,\ ft}$: M+I	**0.219**	0.396	0.743	0.823	0.471	0.590	0.474	**0.582**	**0.811**
9	TC(DN+ChEmb+Bilstm)$^{e=191}$: M+I+Ch+C14+V_{all}	0.233	0.443	0.715	0.804	0.450	0.566	-	-	-
10	TC(ChEmb+Bilstm)$^{e=74,ft}$: M+I+Ch	0.230	0.497	**0.744**	0.804	0.463	0.562	-	-	-
11	TC(DN+ChEmb+Bilstm)$^{e=180,ft}$: M+I+Ch	0.234	**0.516**	0.740	0.813	0.465	0.573	-	-	-
12	TC(DN+ChEmb+Bilstm)$^{e=568,ft}$: M+I+Ch+C14+V_{all}	**0.239**	0.507	0.741	**0.817**	**0.466**	**0.579**	-	-	-
13	TC Ensemble	**0.310**	**0.603**	-	-	-	-	-	-	

obtain moderate improvements, suggesting that transferring knowledge between datasets leading to greater generalization is not trivial.

3 Conclusions and future work

We have presented experimental results of a novel yet simple two-stage conditional chest X-ray report generation model that relies on a topic mining preprocessing of ground-truth reports via regular expressions. Overall, we achieved a new state-of-the-art in several metrics based on the CheXpert labeler, a tool known to better capture the clinical quality of generated reports. Going forward, we see multiple avenues to improve upon this work: (1) improve topic mining, (2) improve topic classification, (3) explore more sophisticated techniques for vision-language multimodal pre-training to increase generalization, (4) train and/or test on other datasets (e.g. Padchest [5]).

Broader Impact

This research attempts to make further progress in the challenging problem of automating the generation of radiology reports by leveraging recent advances in deep learning techniques. From a technical standpoint, we hope to inspire other researchers in the field to explore similar divide-and-conquer approaches to radiology report generation. For example, a possible extension of this work is to use Visual Question Answering (VQA), where instead of prompting the model with topics we could ask the model questions and construct a report from its answers. This could even evolve into a full dialog system that radiologists could interact with, making the system more explainable [15]. We also hope to reinforce the importance of evaluating the medical accuracy of methods, which is why we based our evaluation on the CheXpert labeler [11], but at the same time we acknowledge its limitations and encourage the development of more robust evaluation tools, as well as performing an expert human evaluation of these systems with physicians whenever possible. From a healthcare standpoint, although far from solving the problem, we hope to contribute in the direction of creating AI-based tools for report generation that can complement and enhance the performance of radiologists, and ultimately have a positive impact on the health of millions of patients worldwide who would benefit from efficient and high quality imaging exams. In this sense, we acknowledge that our work lacks a thorough human evaluation with radiologists, and important aspects such explainability, transparency, interpretability, and accountability have to be taken into account and evaluated as well [1, 28, 27].

Acknowledgments and Disclosure of Funding

This work has been funded by Millenium Science Initiative Program ICN2021_004 (iHEALTH), the National Center for Artificial Intelligence CENIA FB210017 and the Center for Mathematical Modeling CMM FB210005, Basal ANID, and the National Agency for Research and Development (ANID) through the Scholarship Program / Doctorado Becas Chile / 2019 - 21191569, and Fondecyt 1210648 and 11201250.

References

[1] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[2] E. Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://aclanthology.org/W19-1909.

[3] Z. Babar, T. van Laarhoven, and E. Marchiori. "Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines". In: *Plos one* 16.11 (2021), e0259639.

[4] W. Boag et al. "Baselines for Chest X-Ray Report Generation". In: *ML4H at NeurIPS*. 2020.

[5] A. Bustos et al. "Padchest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical image analysis* 66 (2020), p. 101797.

[6] Z. Chen et al. "Cross-modal Memory Networks for Radiology Report Generation". 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5904–5914. DOI: 10.18653/v1/2021.acl-long.459. URL: https://aclanthology.org/2021.acl-long.459.

[7] Z. Chen et al. "Generating Radiology Reports via Memory-driven Transformer". In: *EMNLP*. 2020. DOI: 10.18653/v1/2020.emnlp-main.112.

[8] D. Demner-Fushman et al. "Preparing a collection of radiology examinations for distribution and retrieval". In: *Journal of the American Medical Informatics Assoc.* (2015), pp. 304–310.

[9] L. A. Hendricks et al. "Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 570–585. DOI: 10.1162/tacl_a_00385. URL: https://aclanthology.org/2021.tacl-1.35.

[10] G. Huang et al. "Densely connected convolutional networks". In: *CVPR*. 2017. DOI: 10.1109/CVPR.2017.243.

[11] J. Irvin et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *AAAI Conf. on Artificial Intelligence*. 2019. DOI: 10.1609/aaai.v33i01.3301590.

[12] A. Johnson et al. *MIMIC-CXR-JPG-chest radiographs with structured labels (version 2.0.0)*. PhysioNet. 2019. DOI: 10.13026/8360-t248.

[13] N. Kaur, A. Mittal, and G. Singh. "Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey". In: *Multimedia Tools and Applications* (2021), pp. 1–31.

[14] M. Kong et al. "TranSQ: Transformer-Based Semantic Query for Medical Report Generation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 610–620.

[15] H. Lakkaraju et al. "Rethinking Explainability as a Dialogue: A Practitioner's Perspective". In: *arXiv preprint arXiv:2202.01875* (2022).

[16] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. 2004, pp. 74–81.

[17] G. Liu et al. "Clinically Accurate Chest X-Ray Report Generation". In: *ML4H*. 2019.

[18] J. Lovelace and B. Mortazavi. "Learning to Generate Clinically Coherent Chest X-Ray Reports". In: *EMNLP*. 2020. DOI: 10.18653/v1/2020.findings-emnlp.110.

[19] P. Messina et al. "A survey on deep learning and explainability for automatic report generation from medical images". In: *ACM Computing Surveys (CSUR)* (2022).

[20] Y. Miura et al. "Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 5288–5304. DOI: 10.18653/v1/2021.naacl-main.416. URL: https://aclanthology.org/2021.naacl-main.416.

[21] H. Q. Nguyen et al. *VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations*. 2020. arXiv: 2012.15029 [eess.IV].

[22] H. Nguyen et al. "Automated Generation of Accurate & Fluent Medical X-ray Reports". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3552–3569. DOI: 10.18653/v1/2021.emnlp-main.288. URL: https://aclanthology.org/2021.emnlp-main.288.

[23] K. Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *ACL*. 2002. DOI: 10.3115/1073083.1073135.

[24] P. Pino et al. "Clinically correct report generation from chest x-rays using templates". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2021, pp. 654–663.

[25] P. Pino et al. "Inspecting state of the art performance and NLP metrics in image-based medical report generation". In: *arXiv preprint arXiv:2011.09257* (2020). In LXAI at NeurIPS 2020.

[26] A. Radford et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[27] M. Reyes et al. "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities". In: *Radiology: Artificial Intelligence* (2020). DOI: 10.1148/ryai.2020190043.

[28] S. Tonekaboni et al. "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use". In: *Proc of the 4th Machine Learning for Healthcare Conference*. Vol. 106. Proc of Machine Learning Research. Ann Arbor, Michigan: PMLR, Sept. 2019, pp. 359–380.

[29] A. Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[30] X. Wang et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[31] Y. Zhang et al. "Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports". In: *Proc of the 58th Annual Meeting of the ACL*. 2020, pp. 5108–5120.