

Clinically Correct Report Generation from Chest X-rays using Templates

Pablo Pino^{1,3}(✉)[0000-0002-6339-2422], Denis Parra^{1,3}[0000-0001-9878-8761],
Cecilia Besa^{2,4}, and Claudio Lagos^{2,4}[0000-0001-5144-0039]

¹ Department of Computer Science, Pontificia Universidad Católica de Chile, Chile
pdpino@uc.cl, dparra@ing.puc.cl

² School of Medicine, Pontificia Universidad Católica de Chile, Chile
{cbesa, crlagos}@uc.cl

³ Millennium Institute Foundational Research on Data, ANID, Chile

⁴ Millennium Nucleus in Cardiovascular Magnetic Resonance, ANID, Chile

Abstract. We address the task of automatically generating a medical report from chest X-rays. Many authors have proposed deep learning models to solve this task, but they focus mainly on improving NLP metrics, such as BLEU and CIDEr, which are not suitable to measure clinical correctness in clinical reports. In this work, we propose CNN-TRG, a Template-based Report Generation model that detects a set of abnormalities and verbalizes them via fixed sentences, which is much simpler than other state-of-the-art NLG methods and achieves better results in medical correctness metrics.

We benchmark our model in the IU X-ray and MIMIC-CXR datasets against naive baselines as well as deep learning-based models, by employing the Chexpert labeler and MIRQI as clinical correctness evaluations, and NLP metrics as secondary evaluation. We also provide further evidence indicating that traditional NLP metrics are not suitable for this task by presenting their lack of robustness in multiple cases. We show that slightly altering a template-based model can increase NLP metrics considerably while maintaining high clinical performance. Our work contributes by a simple but effective approach for chest X-ray report generation, as well as by supporting a model evaluation focused primarily on clinical correctness metrics and secondarily on NLP metrics.

Keywords: Image report generation · Deep learning · Templates

1 Introduction

Writing a report from medical image studies is an important daily activity for radiologists, yet it is a time-consuming and error-prone task, even for experienced radiologists. AI could alleviate this workload on physicians by providing computer-aided diagnosis (CAD) systems that can analyze an imaging study and generate a written report, which could be used as a starting point by a radiologist to iterate until producing a final report. For chest X-rays, typically, the radiologists examine one or more images from a patient, indicate if there are

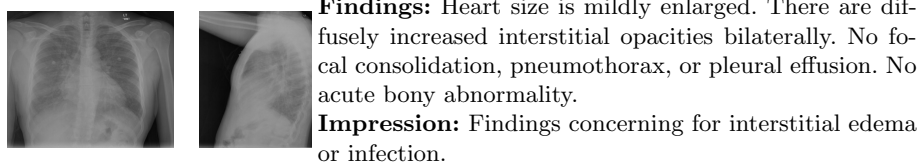


Fig. 1. Imaging study example from the IU X-ray dataset

abnormalities, describe their visual characteristics, and provide a diagnostic or conclusion. Figure 1 provides an example from the IU X-ray dataset [4].

Many deep learning models are proposed in the literature to generate written reports from one or more images [1,2,3,9,10,13,14,17,20,27] employing encoder-decoder architectures or using an image encoder followed by a retrieval or paraphrasing approach. However, it is hard to tell how ready these approaches are for regular clinical use since they are traditionally evaluated by Natural Language Processing (NLP) metrics, such as BLEU [21] or CIDEr-D [28], and these may not be suitable to measure correctness in the medical domain [2,16,17,22,27]. For instance, research on the BLEU metric supports its use for evaluating Machine Translation (MT), but not for other tasks [18,24]. To overcome this problem, some authors have used metrics to evaluate the clinical correctness of the generated reports, such as Chexpert labeler [8] and MIRQI [31], although these have not been tested with expert clinicians nor widely used yet. Moreover, there is a lack of studies on explainability of these systems. This is a highly relevant aspect, since the decisions made from the system predictions will have a direct impact on patients in a clinical setting [25].

In this work, we address the task of report generation from chest X-rays and we make two main contributions. First, we propose CNN-TRG, a deep learning model that detects the presence or absence of abnormalities and then generates the report by relying on a set of pre-defined templates, achieving better performance than state-of-the-art methods in terms of clinical correctness. We also design our model to be simpler and more transparent than the typical encoder-decoder approaches, allowing more control in terms of interpretation. Second, we provide evidence that some traditional NLP metrics are not suitable for evaluating this task by showing they are not robust to textual changes in the reports. Thus, we show that our model can improve its performance in these metrics without affecting clinical correctness.

2 Background and Related Work

Data: Report Structure and Content. Literature [19] shows that the two main datasets used in this task are IU X-ray [4] and MIMIC-CXR [12]. Both contain chest X-rays and their reports written by radiologists, which have two sections of interest: *findings* and *impression*. In *findings*, the radiologist indicates the presence or absence of abnormalities and describes visual characteristics of the positive findings, such as location and severity, among others. In *impression*,

Table 1. Example of a ground-truth *findings* sections and three generated reports, with BLEU (B), ROUGE-L (R-L) and chexpert metrics calculated. Correct and incorrect sentences are **bold** and *italics*, respectively.

Report	NLP		Chexpert		
	B	R-L	F-1	P	R
Ground-truth: Heart size is mildly enlarged. Small right pneumothorax is seen.	-	-	-	-	-
<i>Heart size is normal. No pneumothorax is seen.</i>	0.493	0.715	0	0	0
The cardiac silhouette is enlarged. <i>No pneumothorax.</i>	0.146	0.464	0.5	0.5	0.5
Mild cardiomegaly. Pneumothorax on right lung.	0.075	0.289	1	1	1

the radiologist summarizes the observations into a diagnostic or conclusion. See the example in the Figure 1.

Authors addressing the task of report generation [19] choose one or both of these sections to be generated automatically. We argue that the main information required to write the *findings* section can be observed directly from an image. On the contrary, writing the *impression* may require additional information, such as analyzing multiple views together (frontal and lateral), checking patient symptoms, comparing with prior imaging exams or using medical knowledge that cannot necessarily be inferred from the images alone. Hence, in this article we start by proposing a method to generate the *findings* section of the reports.

Metrics and Clinical Correctness. Most works evaluate the report generation performance using NLP metrics, such as BLEU [21], CIDEr-D, [28] and ROUGE-L [15], which measure n-gram matching between the ground truth and a generated text. These metrics are very popular in machine translation and other NLP tasks; however, they may not be suitable to measure correctness in clinical reports [2,16,17,22,27] or in other tasks [18,24]. To overcome this, some authors have used other metrics to measure the medical accuracy of generated reports. In six previous works [2,3,13,16,17,20], the authors employed the Chexpert labeler [8], a rule-based tool that detects a set of 13 abnormalities from the generated and ground truth reports, and then evaluated these findings using classification metrics. Similarly, Zhang et al. [31] proposed MIRQI, which labels 20 abnormalities and captures visual characteristics described (location, size, etc.). Some authors [1,7,9,30] have used other methods for correctness evaluation, but they do not provide an implementation. To the best of our knowledge, none of these metrics have been validated with expert clinicians, but they aim at clinical accuracy, unlike NLP metrics.

Consider the examples from Table 1, showing a ground truth example, three generated reports, and the performance they achieve using some of these metrics. A generated sample can be clinically incorrect and achieve high NLP scores, or be correct and achieve low NLP scores. Thus, we emphasize the importance of

clinical correctness metrics in this work using the Chexpert labeler and MIRQI as primary metrics above traditional NLP metrics.

Models. The most common approach in the literature derives from the general domain image captioning task with encoder-decoder architectures. Most works use common Convolutional Neural Networks (CNNs) as encoder (e.g. Densenet [6]), and as decoder: a single LSTM to generate word by word [2]; two LSTMs arranged hierarchically to generate sentences and words [9,10,16,31]; or a Transformer-based network [3,17,29]. Some authors [1,13,14,20,27] have employed retrieval or hybrid retrieval-generation approaches for text generation. Compared to other methods, our proposed model CNN-TRG is much simpler, as it uses fewer templates and a more straightforward retrieval process; we test it more thoroughly in the two main datasets available and primarily using medical correctness metrics; and is able to achieve much higher clinical performance.

3 Template-Based Report Generation: CNN-TRG

We propose CNN-TRG, a template-based model that detects abnormalities in the image using a CNN as a classifier and relies on fixed sentences as templates for the text generation. To detect abnormalities, we implement a CNN that receives a chest X-ray and performs multi-label classification of the presence of 13 abnormalities (the chexpert set of labels except for “*No Finding*”). We use Densenet-121 [6], which has shown good results in report generation [19] as well as other medical-related tasks [23]. We trained it using a binary cross-entropy loss for 40 epochs, and applied early stopping by optimizing the PR-AUC classification metric. We initialized the network with the pre-trained weights from ImageNet [5], then trained on the Chexpert dataset [8] for the same classification task, and lastly fine-tuned in the target dataset (IU X-ray or MIMIC-CXR). We used the PyTorch framework⁵ in our implementation⁶.

For text generation, we manually curated a set of two sentences per abnormality indicating presence and absence, totaling 26 sentences. We built the templates by examining the reports and picking existing sentences or creating new ones. To generate the full report, the image is fed to the CNN to compute the binary classification, then the corresponding absence or presence template is chosen for each abnormality, and the sentences are concatenated into the final report. Figure 2 shows the process, and the full details follow in the supplementary material.

We tested the model using two template sets: *single* and *grouped*. Both provide the same meaning clinically (in terms of the presence of the 13 abnormalities), but are written differently.

Single. Concise sentences that indicate the presence or absence directly, for example: “*No pleural effusion*” and “*Pleural effusion is seen*”. The presence

⁵ <https://pytorch.org/>

⁶ <https://pdpino.github.io/clinically-correct>

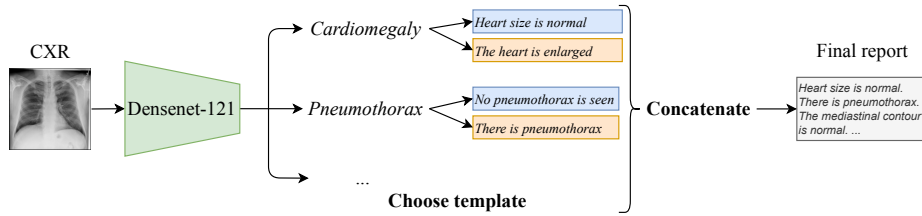


Fig. 2. CNN-TRG model for report-generation

templates do not provide detailed visual characteristics (location, size, etc), since the classification model does not predict this information.

Grouped. To resemble more the reports from each dataset, we grouped multiple abnormalities into common sentences from the training set. For example, in IU X-ray, if all the lung-related abnormalities are classified as absent, the template chosen is “*The lungs are clear*”, instead of using their individual absence templates. If at least one of the abnormalities does not match the group, the model falls back to the *single* set of individual sentences.

4 Experiments

4.1 Datasets

We perform the experiments with two publicly available datasets: IU X-ray⁷ [4] and MIMIC-CXR⁸ [11,12]. Both contain frontal and lateral chest X-rays, IU X-ray has 7,470 images and 3,955 reports, whilst MIMIC-CXR has 377,110 images and 227,827 reports. We used the official train-validation-test split for MIMIC-CXR, and we split the IU X-ray dataset in 80%-10%-10% proportions. We used the *findings* section of the reports and kept only frontal X-rays, leaving a total of 3,311 images in IU X-ray and 243,326 in MIMIC-CXR. To train the CNN in the classification task, we used the chexpert labels provided in MIMIC-CXR, and computed them for IU X-ray applying the Chexpert labeler [8] to the reports.

4.2 Metrics

We used three NLP metrics: BLEU (denoted as B, calculated as the average of BLEU 1-4), ROUGE-L (R-L), and CIDEr-D (C-D), implemented in a publicly available python library⁹. CIDEr-D ranges from 0 (worst) to 10 (best), while the others from 0 (worst) to 1 (best). As clinical correctness metrics, we used Chexpert labeler¹⁰ [8] and MIRQI¹¹ [31], which were detailed in section 2. In

⁷ <https://openi.nlm.nih.gov/faq>

⁸ <https://physionet.org/content/mimic-cxr-jpg/2.0.0/>

⁹ <https://github.com/salaniz/pycocoevalcap>

¹⁰ <https://github.com/stanfordmlgroup/chexpert-labeler>

¹¹ <https://github.com/xiaosongwang/MIRQI>

both cases, we provide F1-score (F-1), precision (P), and recall (R). The chexpert values are the macro average across the 14 labels.

4.3 Baselines

Naive Models. We implement three simple baselines that are not clinically useful, but provide a reference value for the metrics. *Constant*: returns the same report for all the images, manually curated using common sentences from the dataset describing a healthy subject. *Random*: returns a random report from the training set. *1-nn (nearest-neighbor)*: returns the report from the nearest image in the training set, using CNN extracted features from the images as feature space. We used the same CNN as the CNN-TRG model.

Encoder-Decoder Model. We use the CNN from CNN-TRG as encoder and a LSTM with a visual attention mechanism as decoder. The model is trained to generate the full report word by word from the input images. We froze the CNN weights during the report-generation training to avoid over-fitting, and applied early stopping by optimizing the chexpert F-1 score in the validation set. For the LSTM, we used a hidden size of 512 and word embeddings of size 100 initialized with the pre-trained RadGlove [32].

Literature Models. We compare our approach with the results from eleven models [1,2,3,10,13,14,16,17,20,27,31]. We re-implemented the CoAtt model [10], and for the rest we show the results from their papers.

5 Results

Table 2 shows a benchmark of our model against all baselines in both datasets, using the test split. We discuss the results next.

Template Sets. As expected, the clinical performance is the same for both *single* and *grouped* sets, since their clinical meaning is unchanged, but the *grouped* set achieves higher NLP performance, particularly in the IU X-ray dataset. Thus, we show that we can improve NLP metrics only by using more common sentences while preserving the clinical correctness in terms of the seen abnormalities.

CNN-TRG Clinical Correctness. Our template-based models outperform all other models in terms of clinical correctness, both in chexpert and MIRQI F-1 scores. Specifically, our model achieves much better performance than (1) the naive methods, showing it surpasses a first lower standard; and (2) the deep learning models, proving our approach to be more effective while simpler. We also present the results in chexpert F-1 score for each disease in the supplementary material, showing that our model surpasses all other models in every abnormality.

NLP vs Clinical Correctness. Naive models achieve higher NLP performance than CNN-TRG and comparable to some literature models, even though they are not clinically useful by design. On the other hand, naive models achieve very low performance on chexpert and MIRQI, whereas the *CNN-LSTM-att*, literature and CNN-TRG models show higher values. This suggests that these

Table 2. Results in IU X-ray and MIMIC-CXR. Chexpert metrics are macro-averaged across labels. ^{f+i} indicates they generated both *findings* and *impression* sections concatenated, while the rest generated *findings* only; * indicates we re-implemented the code; ^{Ab} indicates they used a subset of the data only with reports that have one or more abnormal findings; super script letters R, T and L indicate Retrieval, Transformer and LSTM-based approaches.

Model	NLP			Chexpert			MIRQI			
	B	R-L	C-D	F-1	P	R	F-1	P	R	
Constant	0.297	0.366	0.307	0.038	0.026	0.071	0.469	0.462	0.481	
Random	0.202	0.284	0.145	0.066	0.065	0.068	0.374	0.378	0.384	
1-nn	0.220	0.301	0.245	0.145	0.150	0.144	0.497	0.508	0.500	
CNN-LSTM-att ^L	0.202	0.319	0.208	0.140	0.159	0.148	0.484	0.492	0.487	
IU X-ray	CoAtt*[10] ^L	0.231	0.316	0.221	0.144	0.162	0.147	0.491	0.503	0.491
	Zhang et al.[31] ^{L,f+i}	0.271	0.367	0.304	-	-	-	0.478	0.490	0.483
	CLARA [1] ^R	0.302	-	0.359	-	-	-	-	-	-
	KERP [14] ^R	0.299	0.339	0.280	-	-	-	-	-	-
	RTEX [13] ^R	-	0.202	-	-	0.193	0.222	-	-	-
	S-M et al.[27] ^{R,f+i}	0.515	0.580	-	-	-	-	-	-	-
	CNN-TRG single	0.167	0.282	0.030	0.239	0.225	0.357	0.529	0.534	0.540
	CNN-TRG grouped	0.273	0.352	0.249	0.239	0.225	0.357	0.529	0.535	0.540
Constant	0.137	0.201	0.059	0.021	0.012	0.071	0.163	0.158	0.176	
Random	0.073	0.142	0.078	0.163	0.186	0.151	0.359	0.372	0.362	
1-nn	0.119	0.193	0.151	0.320	0.325	0.319	0.635	0.645	0.641	
CNN-LSTM-att ^L	0.103	0.244	0.479	0.308	0.378	0.297	0.644	0.652	0.648	
MIMIC-CXR	CoAtt*[10] ^L	0.120	0.252	0.401	0.201	0.356	0.198	0.544	0.551	0.545
	Boag et al. [2] ^L	0.184	-	0.850	0.186	0.304	-	-	-	-
	Liu et al. [16] ^L	0.192	0.306	1.046	-	0.309	0.134	-	-	-
	Chen et al. [3] ^T	0.205	0.277	-	0.276	0.333	0.273	-	-	-
	Lovelace et al. [17] ^T	0.257	0.318	0.316	0.228	0.333	0.217	-	-	-
	CVSE [20] ^{R,Ab}	-	0.153	-	0.253	0.317	0.224	-	-	-
	RTEX [13] ^R	-	0.205	-	-	0.229	0.284	-	-	-
	CNN-TRG single	0.080	0.151	0.026	0.428	0.381	0.531	0.668	0.749	0.640
CNN-TRG grouped	0.094	0.185	0.238	0.428	0.381	0.531	0.666	0.746	0.637	

clinical correctness metrics are better to differentiate automated systems than NLP metrics.

Model Transparency. An advantage of the CNN-TRG model over an end-to-end deep learning approach is the increased transparency. For example, the CNN-LSTM-att baseline performs abnormality detection and text generation inside a black box, while our template-based uses a fully transparent text generation process. Furthermore, by design, our method allows providing a local explanation for each disease independently. Consider Figure 3 showing an input image, the ground truth and generated report, and a Grad-CAM [26] heatmap indicating feature importance for *Cardiomegaly*, the only abnormality found.


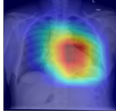
CXR	Generated by CNN-TRG	Ground Truth
 	<p>The heart is enlarged. The mediastinal contour is normal. No focal consolidation. The lungs are free of focal airspace disease. No atelectasis. No pleural effusion. No fibrosis. No pneumonia. No pneumothorax is seen. No pulmonary edema. No pulmonary nodules or mass lesions identified. No fracture is seen.</p>	<p>The heart is mildly enlarged. Left hemidiaphragm is elevated. There is no acute infiltrate or pleural effusion. The mediastinum is unremarkable.</p>

Fig. 3. Example of a report generated with the CNN-TRG using the *single* set of templates, and a Grad-CAM heatmap indicating the activations for the *Cardiomegaly* classification. The colors indicate correct sentences. Best viewed in color.

6 Limitations

The main limitation of our work is that we mostly report the results presented in the original articles. The comparison with other methods could then be improved, since most articles do not provide clinical correctness metrics, and the evaluation protocols may vary. Only MIMIC-CXR has an official train-test split, so the IU X-ray dataset could be more affected by this problem. Additionally, both our templates and the chexpert metric are limited by the set of 13 abnormalities, disregarding their visual characteristics and other chest pathologies. Lastly, our templates are specific to chest X-ray datasets. Hence, in order to use our method with other image modalities or body parts, we would have to manually curate a set of templates covering relevant abnormalities.

7 Conclusions and Future Work

We address the task of automatically generating a text report from chest X-rays and establish a new state-of-the-art in terms of clinical correctness. We present report examples and naive models which challenge the reliability of some traditional NLP metrics to measure model performance, suggesting that text similarity measures might not be suitable in this task. We believe this field should shift to favor clinical correctness instead of traditional NLP metrics to evaluate the systems more appropriately.

As future work, we will replicate implementations from some papers to evaluate and compare their performance under the same experimental conditions. Additionally, we will improve the template-based model by detecting more abnormalities and their visual characteristics, such as location, severity, and more, to provide a more detailed description. We will leverage the templates available at the Radiological Society of North America website¹². Lastly, we will further study the clinical correctness evaluation problem by studying the existing metrics, proposing new ones, and validating them with expert radiologists.

¹² <https://radreport.org/>

Acknowledgments. This work was partially funded by ANID, Millennium Science Initiative Program, Code ICN17.002 and by ANID, Fondecyt grant 1191791.

References

1. Biswal, S., Xiao, C., Glass, L.M., Westover, B., Sun, J.: Clara: Clinical report auto-completion. In: The Web Conf. (2020). <https://doi.org/10.1145/3366423.3380137>
2. Boag, W., Hsu, T.M.H., Mcdermott, M., Berner, G., Alesentzer, E., Szolovits, P.: Baselines for Chest X-Ray Report Generation. In: ML4H at NeurIPS (2020)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: EMNLP (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.112>
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. JAMIA (2015). <https://doi.org/10.1093/jamia/ocv080>
5. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: CVPR (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017). <https://doi.org/10.1109/CVPR.2017.243>
7. Huang, X., Yan, F., Xu, W., Li, M.: Multi-attention and incorporating background information model for chest x-ray image report generation. IEEE Access (2019). <https://doi.org/10.1109/ACCESS.2019.2947134>
8. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI Conf. on Artificial Intelligence (2019). <https://doi.org/10.1609/aaai.v33i01.3301590>
9. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In: ACL (2019). <https://doi.org/10.18653/v1/P19-1657>
10. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: ACL (2018). <https://doi.org/10.18653/v1/P18-1240>
11. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr-jpg-chest radiographs with structured labels (version 2.0.0). PhysioNet (2019). <https://doi.org/10.13026/8360-t248>
12. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data (2019). <https://doi.org/10.1038/s41597-019-0322-0>
13. Kougia, V., Pavlopoulos, J., Papapetrou, P., Gordon, M.: RTECH: A novel framework for ranking, tagging, and explanatory diagnostic captioning of radiography exams. JAMIA (04 2021). <https://doi.org/10.1093/jamia/ocab046>
14. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: AAAI Conf. on Artificial Intelligence (2019). <https://doi.org/10.1609/aaai.v33i01.33016666>
15. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)

16. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: ML4H (2019)
17. Lovelace, J., Mortazavi, B.: Learning to generate clinically coherent chest X-ray reports. In: EMNLP (2020). <https://doi.org/10.18653/v1/2020.findings-emnlp.110>
18. Mathur, N., Baldwin, T., Cohn, T.: Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In: ACL (2020). <https://doi.org/10.18653/v1/2020.acl-main.448>
19. Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., andía, M., Tejos, C., Prieto, C., Capurro, D.: A survey on deep learning and explainability for automatic image-based medical report generation (2020)
20. Ni, J., Hsu, C.N., Gentili, A., McAuley, J.: Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. In: EMNLP (2020). <https://doi.org/10.18653/v1/2020.findings-emnlp.176>
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002). <https://doi.org/10.3115/1073083.1073135>
22. Pino, P., Parra, D., Messina, P., Besa, C., Uribe, S.: Inspecting state of the art performance and NLP metrics in image-based medical report generation. arXiv preprint arXiv:2011.09257 (2020), In LXAI at NeurIPS 2020
23. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning (2017)
24. Reiter, E.: A structured review of the validity of bleu. Computational Linguistics (2018). <https://doi.org/10.1162/coli.a.00322>
25. Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R.: On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiology: Artificial Intelligence (2020). <https://doi.org/10.1148/ryai.2020190043>
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017). <https://doi.org/10.1109/ICCV.2017.74>
27. Syeda-Mahmood, T., Wong, K.C., Gur, Y., Wu, J.T., Jadhav, A., Kashyap, S., Karargyris, A., Pillai, A., Sharma, A., Syed, A.B., et al.: Chest X-ray report generation through fine-grained label learning. In: MICCAI (2020). https://doi.org/10.1007/978-3-030-59713-9_54
28. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015). <https://doi.org/10.1109/CVPR.2015.7299087>
29. Xiong, Y., Du, B., Yan, P.: Reinforced transformer for medical image captioning. In: MLMI (2019). https://doi.org/10.1007/978-3-030-32692-0_77
30. Xue, Y., Xu, T., Long, L.R., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multi-modal recurrent model with attention for automated radiology report generation. In: MICCAI (2018). https://doi.org/10.1007/978-3-030-00928-1_52
31. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. AAAI Conf. on Artificial Intelligence (2020). <https://doi.org/10.1609/aaai.v34i07.6989>
32. Zhang, Y., Ding, D.Y., Qian, T., Manning, C.D., Langlotz, C.P.: Learning to summarize radiology findings. In: LOUHI at NeurIPS (2018). <https://doi.org/10.18653/v1/W18-5623>

8 Supplementary Material

8.1 Template-Based Model

CNN. We used the pytorch implementation¹³ of the Densenet-121 [6] architecture. Specifically, given an input image, we (1) use the `features` layer to extract a feature vector of size $1024 \times H \times W$, (2) apply global average pooling to obtain a vector of size 1024, (3) apply a dropout layer with $p = 0.5$, (4) pass through a fully connected layer to obtain a vector of size 13 with predicted values, and (5) apply a threshold to obtain a binary classification for each abnormality. The specific threshold value for each label is calculated by finding a value that optimizes the F1-score obtained in the validation set.

When training, the weights from the convolutional layers were initialized with ImageNet pre-trained weights from pytorch, and the full model (convolutional and fully connected layer) were pre-trained in the Chexpert dataset. When pre-training in Chexpert, we used a batch size of 54, trained with the Adam optimizer for 15 epochs with learning rate 0.0001 and weight decay (L2-norm) 0.00001. When training in the target dataset (IU X-ray or MIMIC-CXR), we used a batch size of 110, trained with the Adam optimizer for 30 epochs with learning rate 0.00003 and weight decay 0.002. In both cases, we resized the input images to 256×256 , and saved the model with the best PR-AUC evaluated in the validation set. We used a GPU Nvidia RTX 3090 and a GPU Nvidia RTX 2080 for the experiments.

Templates. Tables 3 and 4 show the sentences used in the *single* and *grouped* templates set, respectively. In the *grouped* set we defined multiple groups of abnormalities. For each group, if all its abnormalities match the target (i.e. are predicted positive or negative), the group template is used. If any of the abnormalities does not match the target, the model falls back to using the sentences from the single set. The order of the sentences in the final report for *single* and *grouped* sets is given by the order of the abnormalities or groups in the tables.

8.2 Datasets Pre-processing

We applied the following steps to pre-process the reports: (1) transform letters to lowercase, (2) tokenize, and (3) fixed typos. To extract the *findings* section of the reports from MIMIC-CXR we used publicly available code released by the same authors¹⁴. We removed a few broken images, made a 80%-10%-10% train-validation-test split for the IU X-ray dataset, and used the official split for MIMIC-CXR. Table 5 shows the amounts of images and vocabulary size for each dataset.

¹³ <https://pytorch.org/vision/stable/models.html>

¹⁴ <https://github.com/MIT-LCP/mimic-cxr>

Table 3. Sentences in the *single* template set

Abnormality	Absence template	Presence template
Cardiomegaly (Card)	Heart size is normal	The heart is enlarged
Enlarged Cardiomed. (EC)	The mediastinal contour is normal	The cardiomediastinal silhouette is enlarged
Consolidation (Cons)	No focal consolidation	There is focal consolidation
Lung Opacity (LO)	The lungs are free of focal airspace disease	One or more airspace opacities are seen
Atelectasis (A)	No atelectasis	Appearance suggest atelectasis
Pleural Effusion (PE)	No pleural effusion	Pleural effusion is seen
Pleural Other (PO)	No fibrosis	Pleural thickening is present
Pneumonia (Pn)	No pneumonia	There is evidence of pneumonia
Pneumothorax (Pt)	No pneumothorax is seen	There is pneumothorax
Edema (E)	No pulmonary edema	Pulmonary edema is seen
Lung Lesion (LL)	No pulmonary nodules or mass lesions identified	There are pulmonary nodules or mass identified
Fracture (F)	No fracture is seen	A fracture is identified
Support Devices (SD)	-	A device is seen

Table 4. Sentences in the *grouped* template sets

	Abnormalities	Target	Template
IU X-ray	Heart related (Card, EC)	Negative	The heart size and mediastinal silhouette are within normal limits.
	Lung related (LL, LO, E, Cons, Pn, A, Pt, PE, PO)	Negative	The lungs are clear.
	Pt, PE, LO	Negative	There is no pneumothorax or pleural effusion. No focal airspace disease.
MIMIC-CXR	All abnormalities	Negative	No acute cardiopulmonary process.
	Card, PE, E, A, SD	Positive	In comparison with the study of xxxx, the monitoring and support devices are unchanged. Continued enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases.
	Card, PE, E, A, SD	SD negative, rest positive	Continued enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases.

8.3 Baseline Models

Constant Models. Table 6 shows the reports used in the Constant models for each dataset, both describing a healthy subject.

Encoder-Decoder Model. We trained the CNN-LSTM-att model for 200 epochs in IU X-ray and 30 epochs in MIMIC-CXR, using input image size of

Table 5. Dataset statistics

Dataset	Images	Train	Val	Test	Broken	Vocab size
IU X-ray	3,311	2,638	336	337	4	1,578
MIMIC-CXR	243,326	237,964	1,959	3,403	7	10,161

Table 6. Reports used in the Constant models

Dataset	Constant report
IU X-ray	The heart is normal in size. The mediastinum is unremarkable. The lungs are clear. There is no pneumothorax or pleural effusion. No focal airspace disease. No pleural effusion or pneumothorax.
MIMIC-CXR	In comparison with the study of xxxx, there is little change and no evidence of acute cardiopulmonary disease. The heart is normal in size. The mediastinum is unremarkable. No pneumonia, vascular congestion, or pleural effusion.

256×256 , using a batch size of 120 and an Adam optimizer with learning rate 0.001. We applied early stopping by calculating each epoch the chexpert F-1 score in the validation set, and saving the model with the best performance.

CoAtt Implementation. We replicated the implementation of the CoAtt model [10]. We resized the input images to 450×450 , trained for 150 epochs in IU X-ray and 20 epochs in MIMIC-CXR, and used an Adam optimizer with the same learning rates from the paper: 0.00001 and 0.0005 for the encoder and decoder, respectively. We used the convolutional layers from the template-based model as feature extractor, and we set size 512 for the LSTM hidden size and word embedding size. For IU X-ray we used 586 MTI tags found in the dataset, and used the 14 chexpert labels as semantic tags in MIMIC-CXR, since the MTI tags are not provided. We applied early stopping optimizing the chexpert F-1 score, same as with the CNN-LSTM-att model.

8.4 Metrics

The chexpert labeler classifies each label as *non-mentioned*, *negative*, *uncertain* or *positive*. We consider the former two as negative, and the latter two as positive predictions, similar to other works [16].

8.5 Results

Chexpert Results by Label. Table 7 shows results in chexpert F-1 metric by label. Our template-based model surpasses all other models in every abnormality (all labels except for “*No Finding*”, which indicates no abnormalities).

NLP Results. Table 8 shows NLP performance with the mean and standard deviation calculated across samples in the test set.

Table 7. Chexpert F-1 results by disease in MIMIC-CXR. CSVE [20] used a subset of the data only with reports that have one or more abnormal findings.

F1 by disease	CNN-LSTM-att	1-nn	Boag et al. (1-nn) [2]	Boag et al. (CNN-RNN-beam) [2]	Lovelace et al. [17]	CVSE [20]	CNN-TRG single
No Finding	0.478	0.382	0.455	0.407	0.541	0.300	0.410
Enlarged Cardiomeastinum	0.061	0.112	0.142	0.134	0.059	0.061	0.245
Cardiomegaly	0.497	0.455	0.445	0.390	0.433	0.555	0.583
Lung Lesion	0.066	0.110	0.062	0.001	0.014	0.148	0.155
Lung Opacity	0.287	0.382	0.417	0.077	0.171	0.345	0.563
Edema	0.555	0.492	0.286	0.271	0.298	0.273	0.617
Consolidation	0.096	0.162	0.085	0.014	0.073	0.151	0.265
Pneumonia	0.237	0.330	0.080	0.030	0.039	0.270	0.433
Atelectasis	0.444	0.428	0.375	0.146	0.322	0.398	0.555
Pneumothorax	0.214	0.199	0.111	0.043	0.098	0.060	0.287
Pleural Effusion	0.670	0.613	0.532	0.473	0.480	0.539	0.733
Pleural Other	0.000	0.109	0.039	0.001	0.009	0.058	0.228
Fracture	0.000	0.065	0.060	0.001	0.000	0.056	0.159
Support Devices	0.707	0.648	0.527	0.613	0.660	0.334	0.766
Macro average	0.308	0.320	0.258	0.186	0.228	0.253	0.428

Table 8. NLP results indicating mean \pm standard deviation across samples in the test set of both datasets.

	Model	BLEU	ROUGE-L	CIDEr-D
IU X-ray	Constant	0.297 \pm 0.103	0.366 \pm 0.101	0.307 \pm 0.401
	Random	0.202 \pm 0.095	0.284 \pm 0.094	0.145 \pm 0.581
	1-nn	0.220 \pm 0.124	0.301 \pm 0.116	0.245 \pm 0.889
	CNN-LSTM-att	0.202 \pm 0.116	0.319 \pm 0.114	0.208 \pm 0.474
	CoAtt [10]	0.231 \pm 0.107	0.316 \pm 0.104	0.221 \pm 0.430
	CNN-TRG single	0.167 \pm 0.060	0.282 \pm 0.069	0.030 \pm 0.100
	CNN-TRG grouped	0.273 \pm 0.112	0.352 \pm 0.107	0.249 \pm 0.368
MIMIC-CXR	Constant	0.137 \pm 0.072	0.201 \pm 0.075	0.059 \pm 0.212
	Random	0.073 \pm 0.082	0.142 \pm 0.107	0.078 \pm 0.639
	1-nn	0.119 \pm 0.102	0.193 \pm 0.130	0.151 \pm 0.802
	CNN-LSTM-att	0.103 \pm 0.187	0.244 \pm 0.201	0.479 \pm 1.745
	CoAtt [10]	0.120 \pm 0.166	0.252 \pm 0.178	0.401 \pm 1.536
	CNN-TRG single	0.080 \pm 0.046	0.151 \pm 0.056	0.026 \pm 0.071
	CNN-TRG grouped	0.094 \pm 0.136	0.185 \pm 0.148	0.238 \pm 1.291