Multimodal Image-Text Matching Improves Retrieval-based Chest X-Ray Report Generation

Jaehwan Jeong[*1] JAEHWANJ@STANFORD.EDU
Katherine Tian[*2] KTIAN@COLLEGE.HARVARD.EDU
Andrew Li[1] ANDREWLI@STANFORD.EDU
Sina Hartung[3] SINAHARTUNG@HMS.HARVARD.EDU
Subathra Adithan[4] SUBATHRA26@GMAIL.COM
Fardad Behzadi[5] FBEHZADI1@BWH.HARVARD.EDU
Juan Calle[6] CALLETORO@UTHSCSA.EDU
David Osayande[5] DOSAYANDE@BWH.HARVARD.EDU
Michael Pohlen[1] POHLEN@STANFORD.EDU
Pranav Rajpurkar[3] PRANAV_RAJPURKAR@HMS.HARVARD.EDU

[1] *Stanford University*
[2] *Harvard University*
[3] *Harvard Medical School*
[4] *Jawaharlal Institute of Postgraduate Medical Education and Research*
[5] *Brigham and Women's Hospital*
[6] *University of Texas Health Science Center at San Antonio*

1. Abstract

An AI-based model that automatically generates radiology reports can dramatically improve both clinical workflows and patient care (Boag et al., 2020; Hartung et al., 2020). Such models can triage cases, lessen radiologist workloads, and reduce delays in diagnosis(Nsengiyumva et al., 2021; Dyer et al., 2021; Annarumma et al., 2019). Additionally, AI tools, whether acting autonomously or in collaboration with radiologists, can improve human accuracy (Seah et al., 2021; Sim et al., 2020).

However, current report generation models are not ready for translation into clinical practice (Jing et al., 2018; Chen et al., 2020). While many prior methods adopt image-captioning models to directly generate a report from an image input, their outputs often contain hallucinated information and self-contradictory claims (Yan et al., 2021; Miura et al., 2021). Another stream of work approaches radiology report generation as an image-text retrieval task, as retrieving a human-written report from a corpus of previous reports can guarantee the report's clinical coherence (Endo et al., 2021). However, prior retrieval-based approaches are nowhere near the theoretical upper bound of a retrieval-based generation, as the models often fail to select the ideal report from the corpus (Yu et al., 2022).

To this end, we propose Contrastive X-Ray REport Match (X-REM), an innovative, retrieval-based radiology report generation method that differs from the current paradigm in two key ways. First, our approach emphasizes multimodality, leveraging a language-image

[*] Contributed equally

model to acquire a joint representation of the image and text, as opposed to separately representing them using two unimodal encoders. Secondly, at the retrieval step, we use an image-text matching score as the main similarity metric. In addition to aligning the image and text embeddings in the unsupervised pre-training phase, we also perform supervised contrastive learning that fine-tunes the model to match image and text with the same clinical label. Image-text matching score that is further tuned on the domain-specific features better captures the complicated interaction between a medical image and a report than the conventional cosine similarity, which primarily relies on the alignment of the latent representations.

The structure of X-REM is as follows. First, X-REM sequentially applies a cosine similarity filter and an image-text matching filter to narrow down the retrieval corpus to top j reports most similar to the input image. Then, X-REM removes reports with repetitive information using a natural language inference gate, concatenating the remaining reports into a single document in the final step.

When evaluated on the pre-processed test split of MIMIC-CXR (Johnson et al., 2019), X-REM shows a noticeable improvement over previous radiology report generation modules, including $\mathcal{M}^2$Trans (Miura et al., 2021), R2Gen (Chen et al., 2020), WCL (Yan et al., 2021), CvT2DistilGPT2 (Nicolson et al., 2022), BLIP (Li et al., 2022), and CXR-RePaiR (Endo et al., 2021) on both clinical metrics (3.054 RadCliQ (Yu et al., 2022), 0.133 RadGraph F_1 (Jain et al., 2021), 0.384 CheXbert vector similarity (Smit et al., 2020)) and natural language metrics (0.287 BERTScore (Zhang et al., 2019)). Additionally, we asked radiologists to provide a line-by-line annotation of error scores (No error: 0, Not actionable: 1, Actionable nonurgent error: 2, Urgent error: 3, or Emergent error: 4) on reports generated by X-REM and the baseline retrieval-based approach (CXR-RePaiR). When we analyzed the maximum error severity (MES), which is the worst error in the report, and the average error severity (AES), X-REM reduced the average MES from 2.38 to 2.00 and the average AES from 2.47 to 1.82 compared to the baseline.

		RadCliQ	RadGraph F_1	CheXbert	BERTScore	BLEU2
1	$\mathcal{M}^2$ Trans*	3.087	0.111	0.268	0.227	**0.087**
2	R2Gen*	3.232	0.057	0.203	0.186	0.059
3	WCL*	3.205	0.068	0.218	0.188	0.064
4	CvT2DistilGPT2*	3.191	0.073	0.264	0.193	0.066
5	BLIP	3.304	0.046	0.309	0.190	0.030
6	CXR-RePaiR*	3.181	0.091	0.379	0.191	0.055
7	**X-REM**	**3.054**	**0.133**	**0.384**	**0.287**	0.084

Table 1: X-REM outperforms previous report generation models on multiple metrics. Results with * are from a previous work by Yu et al. (2022) who evaluated the models on the identically-preprocessed MIMIC-CXR test set.

	MES				AES			
	0	≤ 1	≤ 2	≤ 3	0	≤ 1	≤ 2	≤ 3
X-REM (N=117)	0.17	0.35	0.48	0.87	0.23	0.46	0.68	0.91
Baseline (N=69)	0.10	0.33	0.45	0.86	0.11	0.34	0.52	0.82
Human Benchmark (N=53)	0.34	0.49	0.64	0.94	0.35	0.56	0.69	0.94

Table 2: Human Evaluation Study Results. In each row, we show the CDF for each report source across MES and AES scoring. For example, under MES column 3, we have the proportion of reports which have a max error severity of 3 or less.

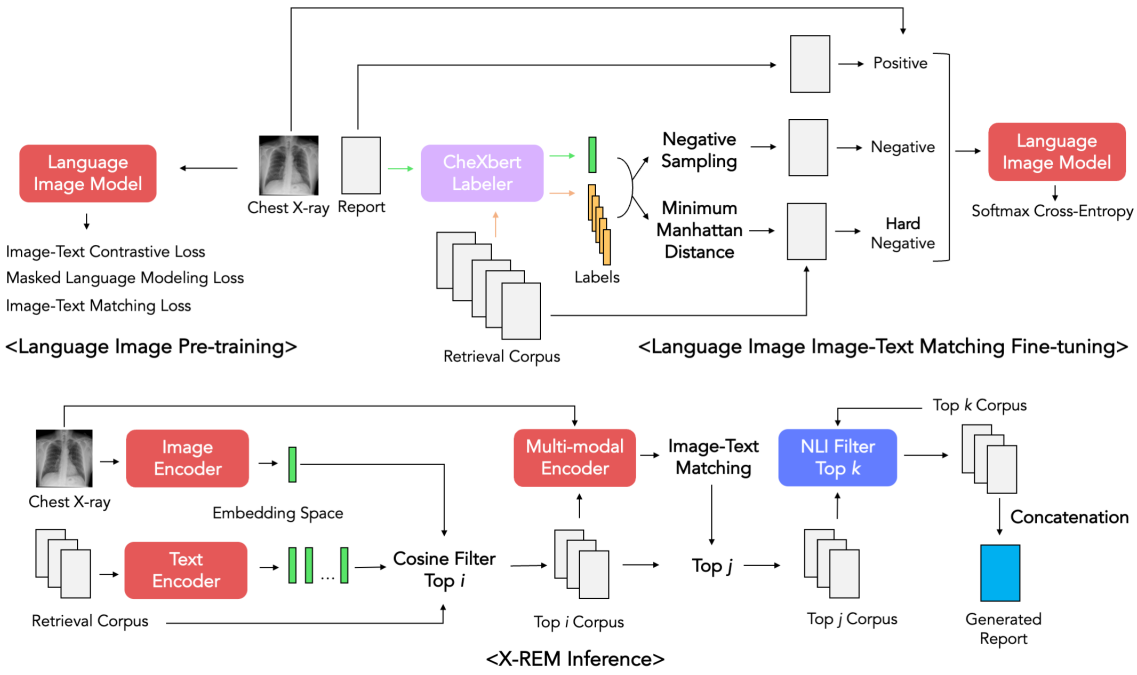

Figure 1: An overview of the pre-training, fine-tuning, and inference steps of X-REM

References

Mauro Annarumma, Samuel J. Withey, Robert J. Bakewell, Emanuele Pesce, Vicky Goh, and Giovanni Montana. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*, 291(1):196–202, 2019. doi: 10.1148/radiol.2018180921. URL https://doi.org/10.1148/radiol.2018180921. PMID: 30667333.

William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for Chest X-Ray Report Generation. In Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 126–140. PMLR, 13 Dec 2020. URL https://proceedings.mlr.press/v116/boag20a.html.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112.

T. Dyer, L. Dillard, M. Harrison, T. Naunton Morgan, R. Tappouni, Q. Malik, and S. Rasalingham. Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm. *Clinical Radiology*, 76(6):473.e9–473.e15, 2021. ISSN 0009-9260. doi: https://doi.org/10.1016/j.crad.2021.01.015. URL https://www.sciencedirect.com/science/article/pii/S0009926021000763.

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 Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Michael P. Hartung, Ian C. Bickle, Frank Gaillard, and Jeffrey P. Kanne. How to create a great radiology report. *RadioGraphics*, 40(6):1658–1670, 2020. doi: 10.1148/rg.2020200020. URL https://doi.org/10.1148/rg.2020200020. PMID: 33001790.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. *CoRR*, abs/2106.14463, 2021. URL https://arxiv.org/abs/2106.14463.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL https://aclanthology.org/P18-1240.

Alistair E. W. 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, 2019. URL https://arxiv.org/abs/1901.07042.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. URL https://arxiv.org/abs/2201.12086.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 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*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm-starting. *CoRR*, abs/2201.09405, 2022. URL https://arxiv.org/abs/2201.09405.

Ntwali Placide Nsengiyumva, Hamidah Hussain, Olivia Oxlade, Arman Majidulla, Ahsana Nazish, Aamir J Khan, Dick Menzies, Faiz Ahmad Khan, and Kevin Schwartzman. Triage of Persons With Tuberculosis Symptoms Using Artificial Intelligence–Based Chest Radiograph Interpretation: A Cost-Effectiveness Analysis. *Open Forum Infectious Diseases*, 8(12), 12 2021. ISSN 2328-8957. doi: 10.1093/ofid/ofab567. URL https://doi.org/10.1093/ofid/ofab567. ofab567.

Jarrel Seah, Cyril Tang, Quinlan Buchlak, Xavier Holt, Jeffrey Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John Lambert, Ben Hachey, Stephen Hogg, Benjamin Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine Jones. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3, 07 2021. doi: 10.1016/S2589-7500(21)00106-0.

Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee, and Byoung Wook Choi. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology*, 294(1):199–209, 2020. doi: 10.1148/radiol.2019182465. URL https://doi.org/10.1148/radiol.2019182465. PMID: 31714194.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.336. URL https://aclanthology.org/2021.findings-emnlp.336.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019. URL http://arxiv.org/abs/1904.09675.