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

Jaehwan Jeong[1], Katherine Tian[2]*, Pranav Rajpurkar[3]*

*[1] Stanford University [2] Harvard University [3] Harvard Medical School * denotes equal contribution*

Overview

- Image-captioning models trained to generate radiology reports from chest X-rays **often output incoherent and incorrect text** due to their lack of medical knowledge
- Retrieval-based report generation **frequently retrieves reports that are irrelevant** to the input X-ray image
- We propose **X-REM**, a retrieval-based radiology report generation model that uses **image-text matching score** to measure the similarity of a chest X-ray image and radiology report for report retrieval
- Image-text matching score with a language-image model can **capture the fine-grained interaction between image and text** that is often lost in cosine similarity
- X-REM **outperforms prior radiology report generation modules** in both natural language and clinical metrics
- Human evaluation of the generated reports suggests that X-REM **increased the number of zero-error reports** and **decreased the average error severity** compared to the baseline retrieval approach

Codebase: github.com/rajpurkarlab/X-REM

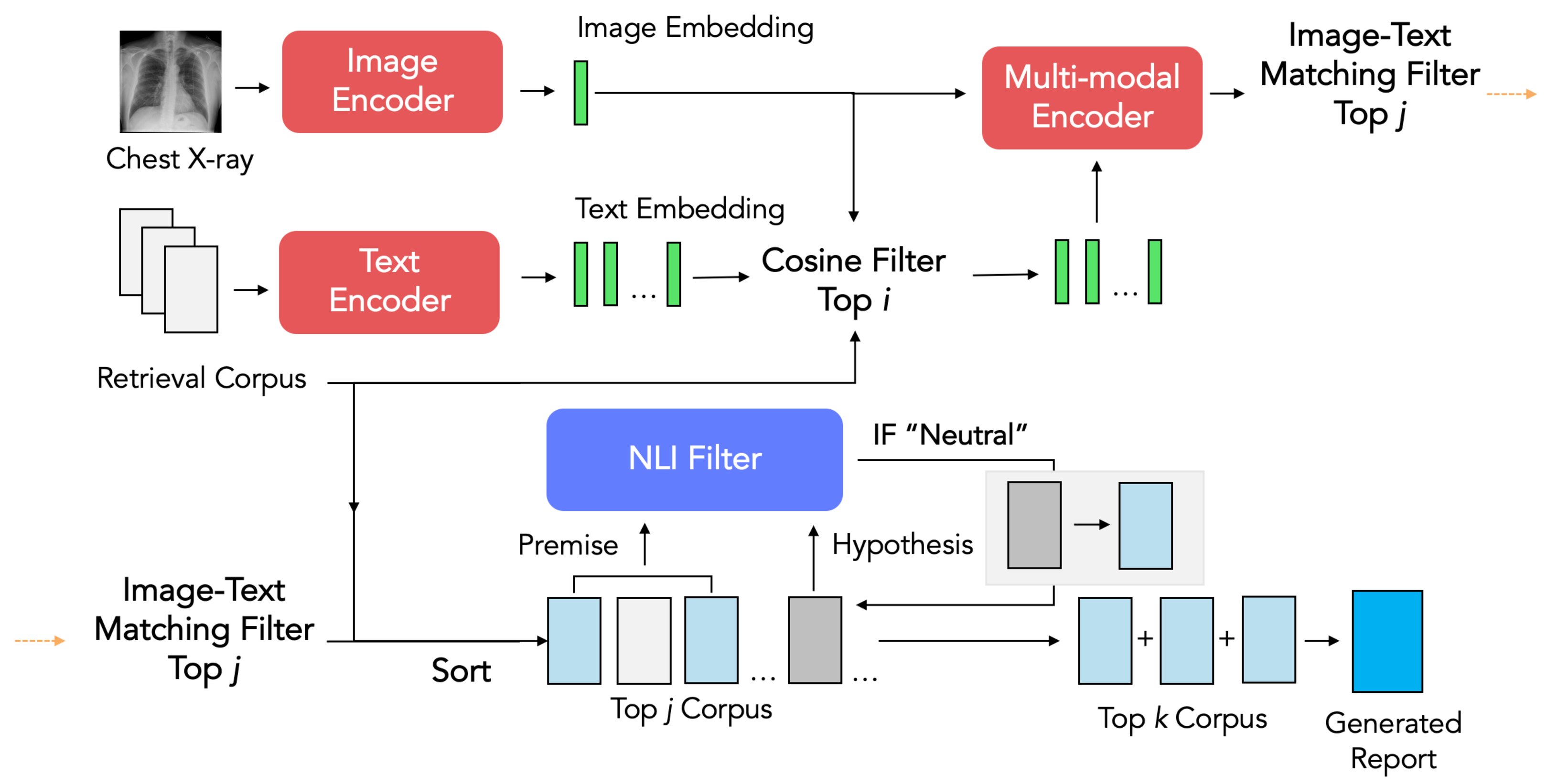
Data and Implementation

- X-REM follows the **architecture and training loss of ALBEF**
  - Architecture: Image Encoder (ViT-B/16), Text Encoder ($BERT_{base}$), Multimodal Encoder ($BERT_{base}$)
  - Training loss: Image-Text Contrastive loss (Pre-training), Masked Language Modeling loss (Pre-training), Image-Text Matching loss (Pre-training and Fine-tuning)
  - X-REM also uses CheXbert for clinical label generation and $BERT_{base}$ tuned on RadNLI/MedNLI for medical NLI
- X-REM is trained on **MIMIC-CXR** to separately generate **impression** and **findings** sections of a radiology report
  - MIMIC-CXR train split is also used as the **retrieval corpus**
- We collaborated with radiologists to conduct a **human evaluation** of the generated reports by **analyzing the clinical errors** in the texts line by line

Methods

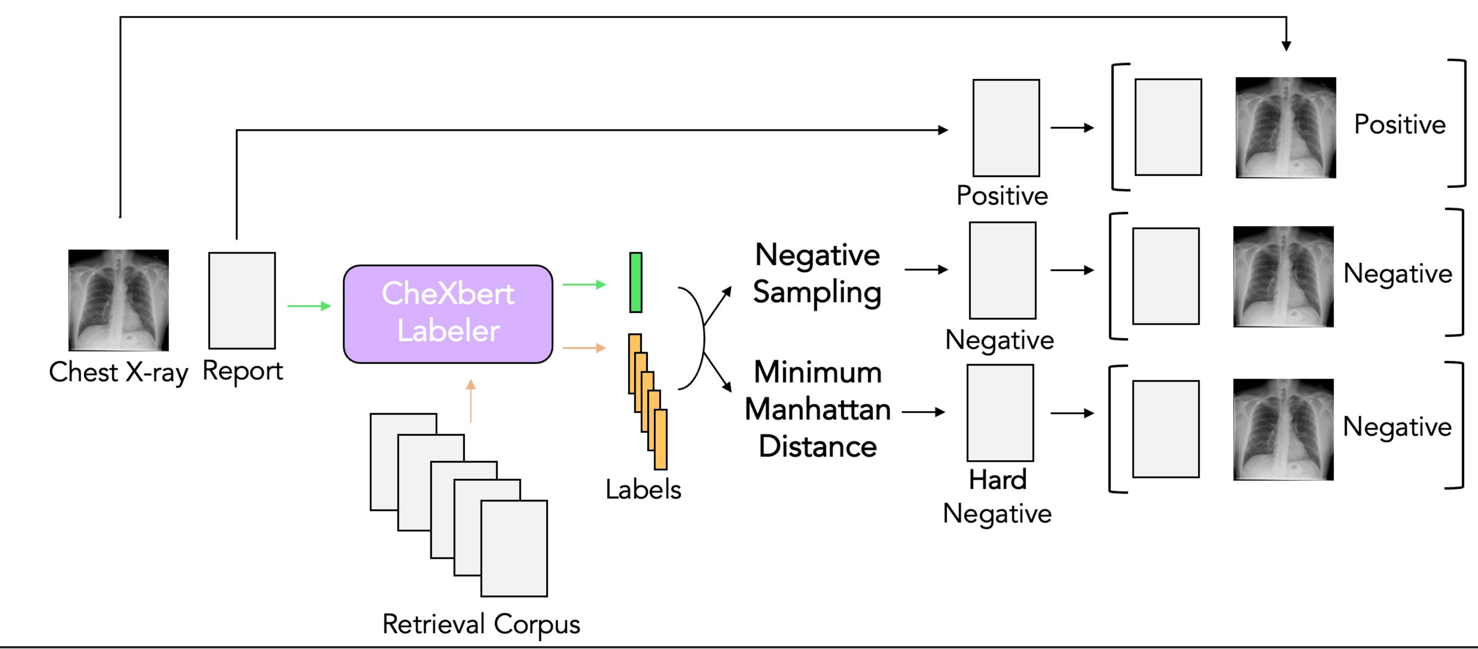
X-REM (Contrastive X-Ray REport Match) Inference

1. Given an input X-ray and a database, retrieve **top *i* reports** that score the highest **cosine similarity**
2. Given top *i* reports, retrieve **top *j* reports** that score the **highest image-text matching (ITM) score**
3. Iterate across the top *j* reports in the decreasing order of ITM scores and filter out **redundant** or **contradictory** reports
4. Concatenate the **top *k* reports** into a single report



Datasest Generation for Image-Text Matching

- X-REM matches studies with **different clinical labels** as **negative samples** for Image-Text Matching
- Studies whose labels have **small non-zero Manhattan distance** serve as **hard negative samples**



Experiments

Results

- X-REM **outperforms** multiple baseline image-captioning models and image-text retrieval models on **RadCliQ**
  - Models were all trained and tested on **MIMIC-CXR**
  - Models with (*) were trained on an additional dataset

	Data	RadCliQ ↓	RadGraph F_1 ↑	CheXbert ↑	BERTScore ↑	BLEU2 ↑
$\mathcal{M}^2$ Trans*	F	**3.277**	**0.244**	**0.452**	**0.386**	**0.220**
X-REM	F	3.585	0.181	0.381	0.353	0.186
CvT2DistilGPT2	F	3.617	0.183	0.375	0.347	0.196
X-REM	I	**3.781**	**0.133**	**0.384**	**0.287**	**0.084**
CXR-RePaiR	I	4.121	0.090	0.379	0.193	0.055
BLIP	I	4.313	0.046	0.309	0.190	0.030
X-REM	I + F	**3.835**	**0.172**	**0.351**	**0.287**	**0.161**
WCL	I + F	3.986	0.143	0.309	0.275	0.144
R2Gen	I + F	4.051	0.134	0.286	0.271	0.137

Human Evaluation

- 5 radiologists each evaluated 60 reports
  - 50% X-REM, 25% CXR-RePaiR, 25% Ground-truth
- Radiologist scored the **clinical error present in each line**
  - No error (0), Not actionable (1), Actionable nonurgent error (2), Urgent error (3), Emergent error (4)
  - Maximum Error Severity is the **maximum of the error scores** in a report
  - Average Error Severity is the **average of the error scores** in a report normalized by the number of lines
- X-REM **outperforms the baseline retrieval method** on both Maximum Error Severity and Average Error Severity

Source	# reports	Maximum Error Severity				Average Error Severity			
		0	≤ 1	≤ 2	≤ 3	0	≤ 1	≤ 2	≤ 3
X-REM	118	0.18	0.36	0.48	0.87	0.24	0.47	0.68	0.91
CXR-RePaiR	69	0.09	0.32	0.45	0.86	0.10	0.33	0.51	0.84
Human Benchmark	53	0.34	0.49	0.64	0.94	0.35	0.56	0.69	0.94

Acknowledgments

We thank Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Subathra Adithan for their work on evaluating radiology reports, and we acknowledge that the project was supported by AWS Promotional Credit