Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model

Mark Endo* MARKENDO@CS.STANFORD.EDU
Stanford University

Rayan Krishnan* RAYANK@CS.STANFORD.EDU
Stanford University

Viswesh Krishna VISWESH@CS.STANFORD.EDU
Stanford University

Andrew Y. Ng ANG@CS.STANFORD.EDU
Stanford University

Pranav Rajpurkar PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Harvard Medical School

Abstract

We propose CXR-RePaiR: a retrieval-based radiology report generation approach using a pre-trained contrastive language-image model. Our method generates clinically accurate reports on both in-distribution and out-of-distribution data. CXR-RePaiR outperforms or matches prior report generation methods on clinical metrics, achieving an average F_1 score of 0.352 ($\Delta + 7.98\%$) on an external radiology dataset (CheXpert). Further, we implement a compression approach used to reduce the size of the reference corpus and speed up the runtime of our retrieval method. With compression, our model maintains similar performance while producing reports 70% faster than the best generative model. Our approach can be broadly useful in improving the diagnostic performance and generalizability of report generation models and enabling their use in clinical workflows.

Keywords: free-text report generation, retrieval, contrastive language-image pre-training

1. Introduction

The automated generation of free-text radiology reports can provide highly interpretable information for doctors, and it has the potential to improve patient care and reduce radiologist workload (Johnson et al., 2019; Chen et al., 2020). Previous methods have shown promising performance in metrics that measure descriptive accuracy (Jing et al., 2018b; Chen et al., 2020), but these systems can fail to produce complete, consistent, and clinically accurate reports (Miura et al., 2021). Developing methods that include relevant, correct, and clear information in reports is critical in order for automated report generation deployment to advance patient care (Hartung et al., 2020; Boag et al., 2020).

In this work, we examine whether we can use self-supervised learning to improve over the best previous methods for report generation. Self-supervised contrastive learning can take advantage of unstructured text data to learn state-of-the-art (SOTA) image representations through the pre-training task of predicting image-text pairs (Radford et al., 2021). This process can be applied to the setting of free-text radiology report generation, where there is an analogous setup of unstructured text and corresponding images.

We develop *Contrastive X-ray-Report Pair Retrieval* (CXR-RePaiR), a retrieval-based radiology report generation method that uses contrastive language image pre-training (CLIP). The novelty of our approach is the use of learned X-ray report pair representations to facilitate the retrieval of unstructured free-text radiology reports. We frame the problem of report generation as a retrieval task instead of an image captioning, language generation task in order to take advantage of highly generalizable zero-shot learning as well as the limited space of possible findings and diagnoses in reports. Specifically, our method uses a CLIP model trained on chest X-

* These authors contributed equally

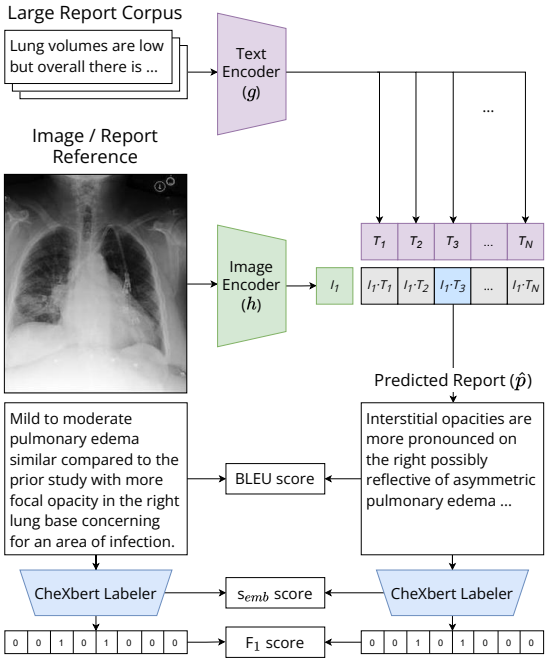

Figure 1: **CXR-RePaiR approach.** Reports or report sentences from a large corpus are passed through a pre-trained text encoder, and the input chest X-ray is similarly passed through a pre-trained image encoder. We generate a prediction by selecting the report that maximizes the similarity between the text and image embeddings. The predicted and ground truth reports are then passed through a labeler and performance scores are computed.

ray-report pairs to rank the similarity between a test dataset of X-rays and a large corpus of reports.

Our method either outperforms or matches the previous SOTA in clinical efficacy (CE) metrics on both the internal dataset (MIMIC-CXR) and an external dataset (CheXpert). In particular, our model has better diagnostic performance over the previous SOTA on the CheXpert dataset with an average F_1 score of 0.352 ($\Delta + 7.98\%$). The benefits of our work over the previous SOTA methods are twofold; (1) we use learned chest X-ray representations that are general and robust to match against a large corpus of reports; (2) our use of retrieval takes advantage of the bounded possibility of described diagnoses. We expect that our method can be broadly useful in improving the diagnostic performance and generability of report generation models and enabling the use of these systems in clinical workflows.

2. Related Work

Previous approaches have structured automated radiology report generation as an image-captioning task (Wang et al., 2018; Yuan et al., 2019). Neural image-to-text radiology report generation models have made use of an image-encoder, text-decoder structure (Wang et al., 2018). Yuan et al. (2019) use a stacked CNN-LSTM structure to produce reports. More recent approaches utilize Transformer decoders (Chen et al., 2020) and attention-based encoders (Cornia et al., 2020).

Recent works by Jing et al. (2018a), Miura et al. (2021), Gale et al. (2018), and Liu et al. (2019) have found that the task of report generation has unique challenges that are not effectively addressed by naive encoder-decoder structures. Radiology reports are longer than typical few sentence summaries produced by image captioning models. Further, the evaluation of generated reports must determine whether disease findings are correctly captured, which may contradict whether the generation uses the same sequence of words as a radiologist (Gale et al., 2018).

New models are being developed to address these unique attributes of radiology report generation (Liu et al., 2019). One technique is to generate reports in a staged text generation procedure (Liu et al., 2019; Nooralahzadeh et al., 2021). In this setup, an image encoder CNN is used to create a hidden representation of the image. Then, a decoder model is used to determine the high-level entities that the report should contain. Finally, a decoder produces the individual sentences from these entities. Yuan et al. (2019) uses a CNN encoder and a hierarchy of LSTM models to extract sentence states and subsequently words from these sentence states. Some works consider more complex methodologies, including a hybrid approach that has a reinforcement-learning trained model that decides when to use a templated report or a generated report (Li et al., 2018), and a graph-neural-network architecture that includes a knowledge graph of disease relationships (Zhang et al., 2020b). While the top-performing prior methods still rely on text generation, we propose a method that addresses the unique properties of report generation by framing the problem as a retrieval task.

Previous works have found that conventional natural language generation (NLG) metrics may be insufficient for the domain of radiology reports. In image captioning and translation, models are evaluated by NLG metrics such as BLEU and METEOR which measure the models' ability to match words and phrases. However, these metrics may not be as useful in the task of report generation. Boag et al. (2020) found that a random retrieval model achieved NLG scores better than an n-gram model, yet performed worse on clinical accuracy. While NLG metrics indicate similarity of vocabulary and phrase structure, the clinical accuracy metric measures the model's ability to produce a report with the correct diagnosis (Boag et al., 2020; Chen et al., 2020; Miura et al., 2021). However, reports contain more valuable information than just the binary labels across pathologies, so Zhang et al. (2020a) introduced the BERTScore metric in which the semantic similarity of sentence pairs is measured. In our work, for completeness, we consider a popular clinical accuracy metric using a state-of-the-art report labeler, a proposed metric to evaluate the semantic similarity, and a traditional NLG metric. We also analyze the relative utility of CE and NLG metrics using example generations.

3. Methods

3.1. Report Generation Using Contrastive Language-Image Pre-training

We define the problem of image-to-text radiology report generation as a retrieval task using a corpus of reports $\mathcal{R} = \{r_1, ..., r_n\}$. For an input chest X-ray x, we generate a predicted report $\hat{p}$ from the corpus $\mathcal{R}$. The predicted report $\hat{p}$ consists of either a report $r \in \mathcal{R}$ or a combination of report sentences $s_1, ..., s_k \in \mathcal{S}(\mathcal{R})$, where $\mathcal{S}(\mathcal{R})$ is the set of all sentences in $\mathcal{R}$. Our method CXR-RePaiR-R returns $\hat{p}$ as the report in $\mathcal{R}$ that maximizes $f(r, x)$, where f is the CLIP model scoring function. Separately, our method CXR-RePaiR-k returns $\hat{p}$ as the top k sentences in $\mathcal{S}(\mathcal{R})$ that maximize $f(s, x)$.

The CLIP model passes r (or s) through a text encoder g to produce a text embedding T. Similarly, it passes x through an image encoder h to produce an image embedding I. We then compute the similarity score $f(r, x) = g(r) \cdot h(x) = T \cdot I$ (or $f(s, x) = g(s) \cdot h(x)$ for sentences). We use a CLIP model first pre-trained on natural image-text pairs and subsequently trained on radiology report-image pairs. In these training phases, the model is trained to produce higher dot-product similarity values for image-text pairs of the same instance, and low dot-product values for pairs of separate instances.

Figure 1 details the report generation and evaluation process.

3.2. CXR-RePaiR Sentence Selection

As we later discuss in Section 6.3, the number of sentences in $\mathcal{S}(\mathcal{R})$ that we want to return as $\hat{p}$ may vary depending on the contents of the predicted sentences. For example, if $\arg\max_{s \in \mathcal{S}(\mathcal{R})} f(s, x)$ is a sentence that describes no finding, then it may be disadvantageous to return a final report with lots of other sentences. To counteract this potential issue, we develop a method CXR-RePaiR-Select that selects for a varying number of sentences to include in the final report. Specifically, we feed $\arg\max_{s \in \mathcal{S}(\mathcal{R})} f(s, x)$ into the CheXbert labeler (Smit et al., 2020), which outputs diagnostic predictions for the sentence. If the prediction includes no finding, then $\hat{p}$ is chosen to be $\arg\max_{s \in \mathcal{S}(\mathcal{R})} f(s, x)$. Otherwise, $\hat{p}$ is the top k sentences that maximize $f(s, x)$.

3.3. Corpus Generation

We consider corpus generation from both reports ($\mathcal{R}$) as well as individual sentences from reports ($\mathcal{S}(\mathcal{R})$). We hypothesize that generating the predicted report $\hat{p}$ by combining sentences in $\mathcal{S}(\mathcal{R})$ rather than using a single full report in $\mathcal{R}$ will result in more effective retrieval as it increases the prediction space.

3.4. Baselines

We compare our method to two report generation studies: R2Gen (Chen et al., 2020) and $\mathcal{M}^2$ Trans[1] (Miura et al., 2021). These methods were selected as the two top performing models, and $\mathcal{M}^2$ Trans is the prior SOTA in average F_1 score.

Retrieval Baseline In order to understand the relative impact of general contrastive self-supervised pre-training, we develop a retrieval baseline that uses image classification model embeddings to score report similarities. Specifically, the baseline model selects a report r in $\mathcal{R}$ based on its corresponding image i. For an input image x, the model returns $\hat{p}$ as the

1. Using $\mathcal{M}^2$ Trans w/ NLL+BS+fc_E configuration.

Table 1: Evaluation of CXR-RePaiR method and baselines on internal (MIMIC-CXR) and external (CheXpert) datasets. Performance is evaluated using our s_{emb} metric, BLEU2 score, and F_1 (macro-avg) score. We find that our approach either outperforms or matches previous SOTA in clinical scores on both datasets.

Dataset	Model	Evaluation Metrics		
		s_{emb}	BL2	F_1
Internal (MIMIC-CXR)	R2Gen (Chen et al., 2020)	0.202 ± 0.002	0.212 ± 0.001	0.148 ± 0.003
	$\mathcal{M}^2$ Trans (Miura et al., 2021)	0.247 ± 0.002	**0.217 ± 0.001**	0.270 ± 0.004
	Retrieval Baseline (Ours)	0.360 ± 0.002	0.092 ± 0.001	0.208 ± 0.003
	CXR-RePaiR-2 (Ours)	**0.385 ± 0.003**	0.069 ± 0.001	0.256 ± 0.004
	CXR-RePaiR-Select (Ours)	0.343 ± 0.004	0.050 ± 0.001	**0.274 ± 0.003**
External (CheXpert)	R2Gen (Chen et al., 2020)	0.269 ± 0.005	0.168 ± 0.004	0.191 ± 0.008
	$\mathcal{M}^2$ Trans (Miura et al., 2021)	0.307 ± 0.006	**0.179 ± 0.003**	0.326 ± 0.015
	Retrieval Baseline (Ours)	0.283 ± 0.005	0.069 ± 0.002	0.246 ± 0.007
	CXR-RePaiR-2 (Ours)	**0.317 ± 0.008**	0.088 ± 0.002	0.310 ± 0.009
	CXR-RePaiR-Select (Ours)	0.292 ± 0.008	0.073 ± 0.001	**0.352 ± 0.005**

report that corresponds the the image which maximizes $\frac{C(x)C(i)}{\|C(x)\|\|C(i)\|}$. Here, C is a ResNet18 encoder (He et al., 2016) trained on CheXpert images (Irvin et al., 2019). This model directly assumes that chest X-rays with similar visual features have similar radiology reports.

3.5. Evaluation

We focus on two metrics to evaluate the clinical efficacy (CE) of the generated reports. The first CE metric is the F_1 (macro-avg) score which evaluates the accuracy of the generated reports' diagnostic predictions. Generated reports are passed through the automated CheXbert labeler which has near radiologist performance in labeling medical conditions (Smit et al., 2020). The F_1 score is then computed for these predictions against the ground truth labels. The average F_1 score is computed from all 14 radiological label categories. For the second CE metric, we introduce a new score to measure the semantic similarity of the generated report compared to the original report. The generated report and true report are passed through the CheXbert labeler, and the cosine similarity between the last hidden representations is calculated to produce s_{emb}. We also include BLEU2 scores as a natural language generation (NLG) metric, even though NLG metrics have been shown to be flawed and inadequate for evaluating free-text radiology reports (Boag et al., 2020; Miura et al., 2021). For our evaluations, we construct bootstrap confidence intervals by creating 10 resamples with replacement from each test dataset, where each resample is the size of the original test set.

4. Data and Implementation

We use the MIMIC-CXR dataset for generating the corpus of radiology reports used during retrieval, pretraining the CLIP model, and training baseline models. MIMIC-CXR v2.0.0 is a large dataset of chest X-rays and associated free-text reports with 180,000 imaging studies and more than 300,000 images taken from 64,588 patients (Johnson et al., 2019). The labels are structured in a multi-label multi-class manner with each of the 14 classes having labels indicating presence, absence, or uncertainty.

We evaluate on both the in-distribution MIMIC-CXR test set and the out-of-distribution CheXpert test set. CheXpert is a dataset of 224,316 chest x-rays from 65,240 patients, with associated labels for 14 observations from radiology reports (Irvin et al., 2019). In particular, the CheXpert test set consists of 500 studies with 688 images and has corresponding radiology reports.

Radiology reports have many sections including examination, indication, technique, findings, and impression. For this research, we choose to predict the impression section as it includes the radiologist's diagnosis for the X-ray. In addition, the findings section

Table 2: Examples of different methods' generated reports compared to the reference report. Correct positive predictions are color coded by finding type to improve readability. On these examples, our method does particularly well in providing clinically useful information that is accurate, clear, and actionable.

Reference	R2Gen	$\mathcal{M}^2$ Trans	CXR-RePaiR-3 (Ours)
Findings : Frontal and lateral chest radiographs demonstrate minimal blunting of the bilateral costophrenic angles. There is no focal consolidation or pneumothorax. The heart size is moderately enlarged and there are post-surgical changes of median sternotomy and CABG. There is an indistinct appearance of the pulmonary vasculature consistent with mild-to-moderate pulmonary edema. **Impression** : Mild-to-moderate pulmonary edema without focal consolidation to suggest pneumonia. There may be small bilateral pleural effusions with a small degree of bilateral atelectasis. The heart is moderately enlarged.	Frontal and lateral views of the chest were obtained. The patient is status post median sternotomy and CABG. There are relatively low lung volumes. There is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema versus atypical infection. No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are stable.	Heart size is normal. The mediastinal and hilar contours are normal. The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Median sternotomy wires are intact.	Cardiomegaly with left pleural effusion and mild edema. Large left pleural effusion with mild pulmonary vascular congestion. The severe enlargement of cardiac silhouette may have progressed the volume of presumed right pleural effusion is impossible to assess on a single frontal view.
Findings : The lungs appear hyperexpanded. There is mild increased pulmonary vascular congestion from ____. A small right pleural effusion is likely present with mild right basilar atelectasis. Right base consolidation is not entirely excluded. No significant left pleural effusion or pneumothorax is detected. Suture chain material and scarring in the left upper-to-mid lung zone is not significantly changed. Multiple mediastinal surgical clips are compatible with history of CABG surgery. The cardiac silhouette is top normal in size but unchanged. The mediastinal and hilar contours are within normal limits with moderate tortuosity of the descending thoracic aorta. Lobulation at the apex of the left hemi thorax along the mediastinal border is stable, residual of slowly resolving hematoma. **Impression** : 1. Increased mild pulmonary vascular congestion from ____ with small right pleural effusion and right basilar atelectasis. Right basilar opacity may be combination of above but underlying consolidation due to infection is not excluded. 2. Staple suture material and scar in the left upper-to-mid lung.	AP portable upright view of the chest. Lung volumes are low limiting assessment. Overlying EKG leads are present. Allowing for this there is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact.	Heart size remains mildly enlarged. The mediastinal and hilar contours are unchanged. There is mild pulmonary edema. There is no focal consolidation. There is no large pleural effusion or pneumothorax.	Pulmonary edema with right base opacity compatible with atelectasis but early infection cannot be excluded. More confluent right basilar opacity may relate to pulmonary edema and small right pleural effusion however underlying consolidation is not excluded. Coarse interstitial markings more prominent in the right lower lung field associated with small bilateral pleural effusions with concurrent bibasilar atelectasis right worse than left findings suggest mild vascular congestion.

atelectasis edema cardiomegaly pleural effusion lung opacity

is often missing from the report (which is particularly important for the relatively small CheXpert test set). Since the generative baseline models predict the findings section, it is used when available as the ground truth report to evaluate the BLEU2 score for these methods, and the impression section is used otherwise.[2]

2. 2834 out of the 3678 evaluated MIMIC-CXR test studies have a findings section in the report.

5. Experiments

5.1. CXR-RePaiR Report Generation

We investigate the performance of a contrastive language image model trained on radiology report-image pairs as a retrieval-based report generation method. We compare our CXR-RePaiR method and our retrieval baseline against previous approaches in Table 1.

We find that our CXR-RePaiR approach produces clinically accurate reports on both the in-distribution MIMIC-CXR test set and the out of distribution CheXpert test set. On CheXpert, CXR-RePaiR-Select with $k = 6$ sentences achieves an F_1 score of 0.352 ($\Delta + 7.98\%$) and CXR-RePaiR-2 achieves an s_{emb} score of 0.317. The next highest-performing highest-performing model is $\mathcal{M}^2$ Trans (Miura et al., 2021) with an F_1 score of 0.326 and s_{emb} score of 0.307. On MIMIC-CXR, CXR-RePaiR-Select has an F_1 score of 0.274, slightly beating $\mathcal{M}^2$ Trans with a score of 0.270. CXR-RePaiR-2 has an s_{emb} score of 0.385, which is 55.9% greater than the SOTA generative method.

We also qualitatively compare some of the generated reports across methods. Table 2 includes reports generated for different models for two test images. We find that our method generally produces clinically accurate descriptions, but it does not always use the same language as the original report. For instance, in the first example, the reference writes that "the heart size is moderately enlarged" while our method generates "enlargement of [the] cardiac silhouette" and "cardiomegaly." These statements are clinically equivalent, but their verbiage differs. In contrast, the $\mathcal{M}^2$ Trans model generates "heart size is normal," which is an incorrect diagnosis despite matching n-grams to the reference. These observed findings are in agreement with Boag et al. (2020) which note the lack of concordance of NLG metrics with generated report accuracy.

Finally, we observe that our baseline retrieval method performs surprisingly well. This may be explained by similar chest x-ray images having similar reports, as well as the success of a retrieval method with access to a large reference corpus.

6. Ablation Studies

We analyze the increase in performance of CXR-RePaiR with (i) contrastive language-image pre-training, (ii) using sentences instead of reports to increase the expression space, and (iii) the increase in the number of sentences k used to create the report.

6.1. Lanugage-Image Model Pre-training Techniques

We study the effect of three contrastive language-image pre-training procedures on the performance of the CXR-RePaiR approach. We expect that contrastive pre-training on natural language-image pairs and subsequently finetuning on radiology report-image pairs will outperform a model solely trained on radiology report-image pairs.

	Corpus	F_1	s_{emb}
CXR-RePaiR-Imp	$\mathcal{R}$	0.200	0.236
CXR-RePaiR-Full	$\mathcal{R}$	0.168	0.222
CXR-RePaiR-Pretrain	$\mathcal{R}$	0.294	0.268
	$\mathcal{S}(\mathcal{R})$	**0.310**	**0.317**

Table 3: Comparison of three CLIP training methods. CXR-RePaiR-Imp is trained on the impressions section alone, CXR-RePaiR-Full is trained on the full reports, and CXR-RePaiR-Pretrain is first pre-trained on natural images and subsequently trained on the impressions section.

We construct three CXR-RePaiR models: CXR-RePaiR-Imp (trained contrastively on radiology report impression section and image pairs), CXR-RePaiR-Full (trained contrastively on the full radiology report and image pairs with an extended context length), and CXR-RePaiR-Pretrain (first pre-trained contrastively on natural language-image pairs and then trained contrastively on radiology report impression section and image pairs). We find that CXR-RePaiR-Pretrain significantly outperforms the other approaches, as expected. CXR-RePaiR-Pretrain has an average F_1 score of 0.294 while CXR-RePaiR-Imp and CXR-RePaiR-Full have average scores of 0.200 and 0.168 respectively. Detailed results can be found in Table 3.

6.2. Sentence vs. Report Retrieval

Since report impressions are comprised of individual sentences, the task of report generation using retrieval can be framed as either (1) the selection of an entire impression that best matches the inputted X-ray or (2) the selection of sentences that individually best match the inputted X-ray. It is possible that the first methodology may perform better because the model was explicitly pre-trained to match entire impressions, rather than individual sentences, with images. However, this procedure is restricted by the limited amount of available reports in the retrieval corpus. In order for this setup to succeed, the findings and diagnosis of a patient in the test set must

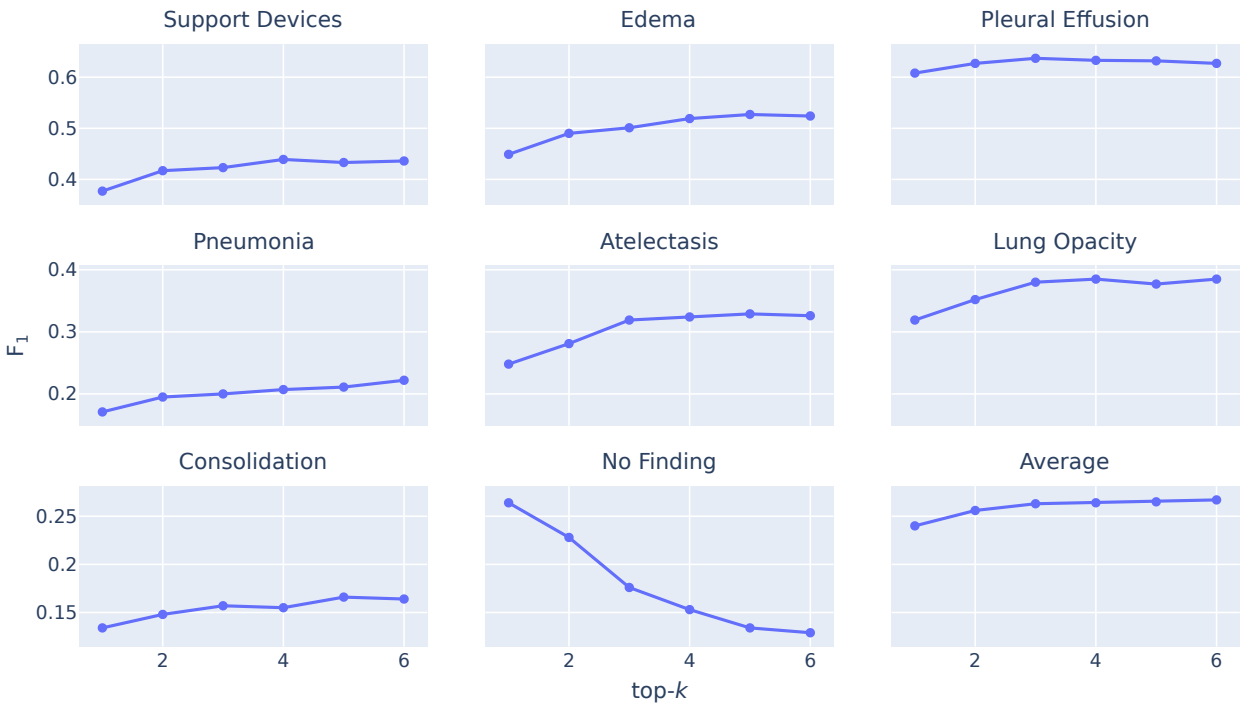

Figure 2: Evaluation of models that produce reports of k sentences across radiological findings on MIMIC-CXR test set. The average F_1 score improves as k increases, but the "no finding" score reduces as k increases.

exactly match that of a previously seen patient in the training set. Therefore, we experiment with selecting individual sentences which may come from different reports. We predict this is a superior method because it has the potential to improve the model expressivity by greatly increasing the space of possible outputs the model can produce. We seek to answer this question of whether our CXR-RePaiR method performs better constructing reports from individual sentences or matching the entire report section directly.

Indeed, we see that using a corpus of report sentences ($\mathcal{S}(\mathcal{R})$) results in improved performance over using a corpus of reports ($\mathcal{R}$). As hypothesized, this is likely attributed to the model's ability to match multiple independent sentences across various reports for a single test example. The report-level model is significantly limited in that if the findings or diagnosis of a patient is not seen identically in the training set, the model will be guaranteed to have an incorrect prediction. On the other hand, the sentence-level model has some ability to make predictions outside of the explicit space of the reference set. Results are shown in Table 3.

6.3. Top-k Sentences

Because we find that generating reports from sentences results in improved performance, we also experiment with different values for the hyperparameter k where k is the number of top candidate sentences selected. When framing report generation as the selection of the top-k best matching sentences, we observe that increasing report length increases the average F_1 score. However, we find that increasing report length does not strictly improve F_1 scores across findings. In particular, the "no finding" F_1 score reduces when increasing the number of sentences. This can be attributed to the property that radiologists generally write brief and short reports to indicate no finding. By requiring more than one sentence to be included, the model is more likely to incorporate an incorrect finding/diagnosis and have a diminished F_1 score. Overall, this suggests that in most cases the

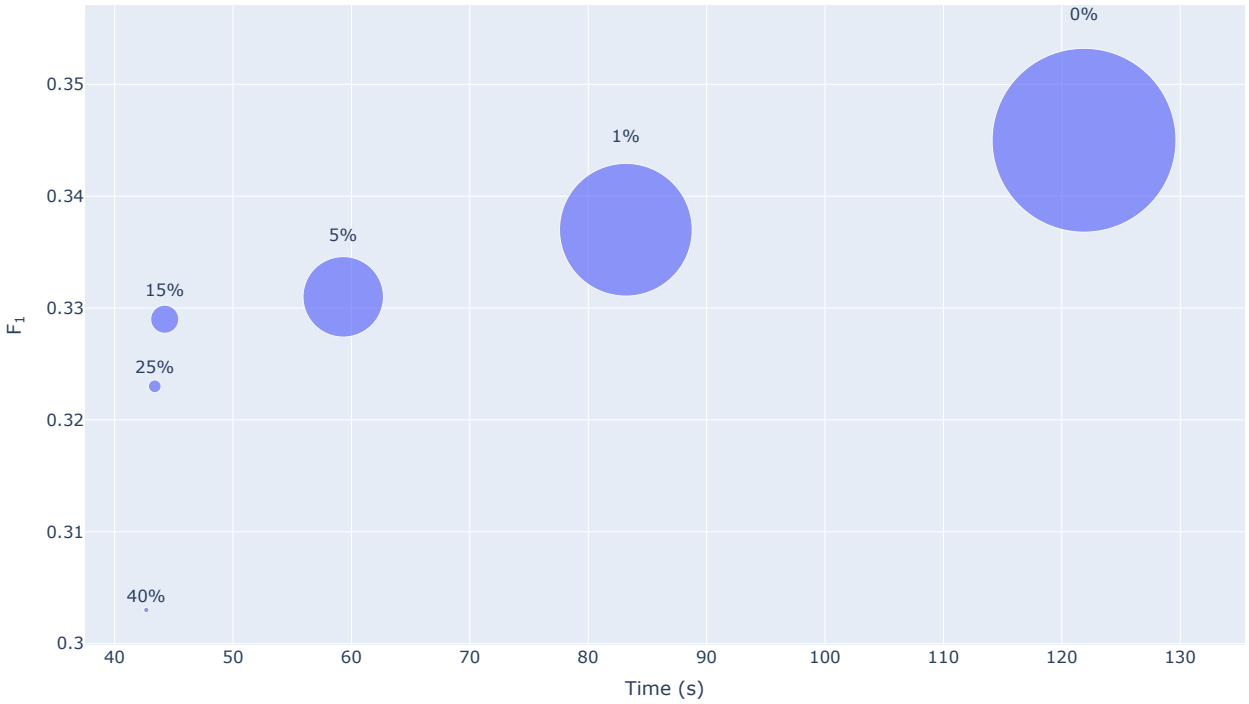

Figure 3: **Runtime vs F_1 analysis for various corpus compression levels.** The label denotes the compression level and the area of the bubble illustrates the compressed corpus size at the corresponding compression level. We observe that small compression levels significantly reduce corpus size and consequently runtime while maintaining clinical accuracy. The model is evaluated on the CheXpert dataset.

prediction quality improves when generating numerous descriptions of possible findings and diagnoses, though a case without the presence of any pathologies benefits from a short description. Figure 2 displays results across 8 chosen radiological categories and average scores.

7. Runtime Considerations

We compress the report corpus to enable faster retrieval. The report corpus is compressed as follows. First, CheXbert embeddings are extracted for every element of the corpus. Second, cosine similarity is used to cluster similar elements using a defined threshold value. Finally, representative elements are extracted from each cluster to form the compressed corpus. We experiment with threshold values of 40%, 25%, 15%, 5%, 1% and 0% (no compression). For testing, we use a single NVIDIA Titan X.

Runtime, F_1 scores, and corpus size for various compression levels are illustrated in Figure 3. We find that compression can significantly reduce runtime and corpus size while maintaining clinical accuracy. Small compression levels reduce corpus size significantly; a 15% compression level reduces corpus size by 40× and reduces runtime by 65% compared to the model without compression while maintaining comparable accuracy. We observe that a 15% compression level provides the best tradeoff between accuracy ($F_1 = 0.329$) and runtime ($t = 42$s).

A comparison of runtime and F_1 scores for CXR-RePaiR-C (15% compression level), CXR-RePaiR-6, and previous approaches can be found in Figure 4. We find that CXR-RePaiR-6 outperforms current models in clinical accuracy while maintaining comparable runtime. CXR-RePaiR-C maintains high clinical accuracy while introducing a 65% runtime speedup.

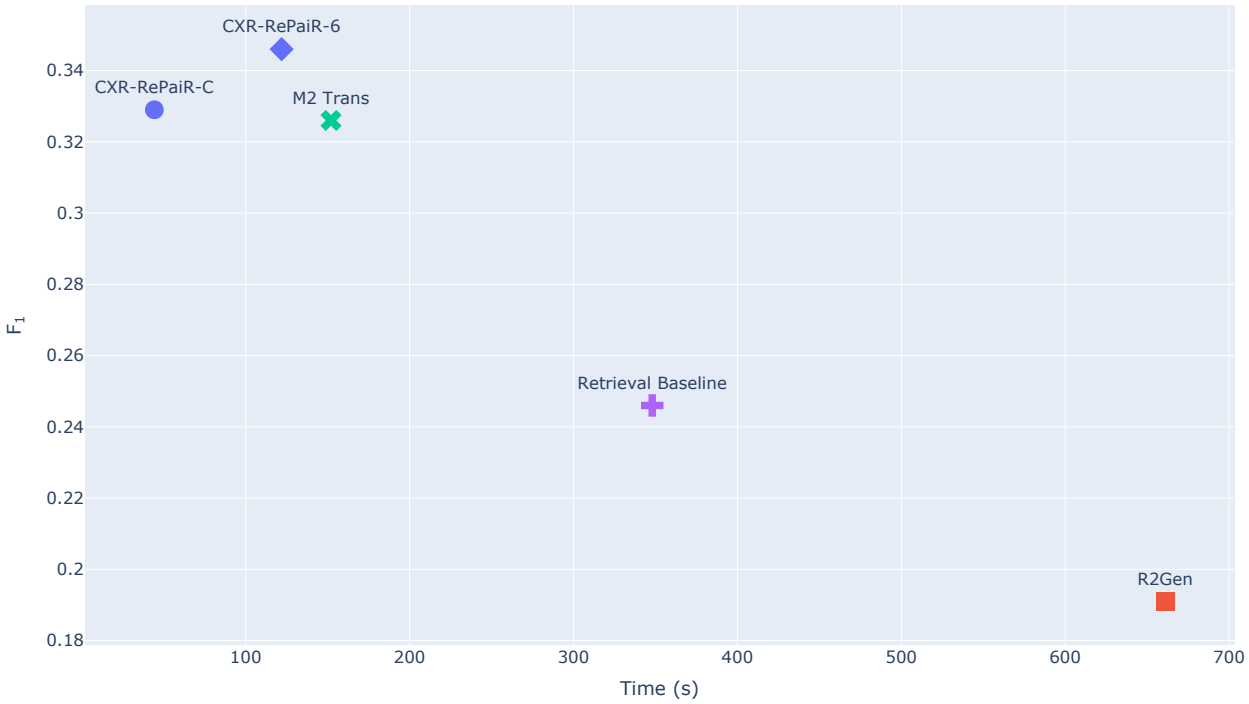

Figure 4: **Runtime vs F_1 comparison between CXR-RePaiR and prior methods.** We compare our CXR-RePaiR models without compression and with a 15% compression level against baselines. Our CXR-RePaiR-6 model (without compression) achieves a high F_1 while maintaining similar runtime to the current state-of-the-art. Our CXR-RePaiR-C model with 15% compression maintains comparable clinical accuracy to our best CXR-RePaiR-6 model while reducing runtime by 65%. Models are evaluated on the CheXpert dataset.

8. Limitations and Future Work

The success of this methodology is a direct consequence of the size and variability of the reference corpus. In particular, the model cannot be expected to make predictions for rare pathologies if they do not appear in the reference corpus. Future work may consider more precise ways to construct an exhaustive reference corpus.

Though our method produces clinically accurate descriptions, we find that the CXR-RePaiR-k setup can be prone to repeating information as the sentence selection process solely depends on maximizing the CLIP scoring function. Future work can incorporate choosing sentences based on the additive value of their diagnostic information. Our method can also be easily applied to report templating, where a crafted set of templates can be used instead of the large corpus of reports.

9. Conclusion

We find that CXR-RePaiR—a retrieval-based chest X-ray report generation method utilizing a pre-trained contrastive language-image model—produces state-of-the-art results on clinical efficacy metrics. In the ablation studies, we describe how choices in pre-training technique, retrieval corpus type, and the number of selected sentences improve model performance. Our best model is pre-trained on natural data and subsequently trained on medical data, and reports are generated as a set of sentences. Lastly, we consider how to effectively reduce the runtime of our model without considerably sacrificing performance. Our approach demonstrates the power of using transferable representation learning along with a large corpus of existing reports to generate clinically accurate, clear, and actionable text that can advance patient care.

References

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.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning, 2020.

William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. Producing radiologist-quality reports for interpretable artificial intelligence, 2018.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018a. doi: 10.18653/v1/p18-1240. URL http://dx.doi.org/10.18653/v1/P18-1240.

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 2018b. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL https://aclanthology.org/P18-1240.

Alistair EW 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. *arXiv preprint arXiv:1901.07042*, 2019.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation, 2018.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR. URL http://proceedings.mlr.press/v106/liu19a.html.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation, 2021.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish

Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, 2018.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020a.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph, 2020b.