Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays

Jianmo Ni[*], **Chun-Nan Hsu, Amilcare Gentili, Julian McAuley**
University of California, San Diego
{jin018,chunnan,agentili,jmcauley}@ucsd.edu

Abstract

Automatic medical image report generation has drawn growing attention due to its potential to alleviate radiologists' workload. Existing work on report generation often trains encoder-decoder networks to generate complete reports. However, such models are affected by data bias (e.g. label imbalance) and face common issues inherent in text generation models (e.g. repetition). In this work, we focus on reporting abnormal findings on radiology images; instead of training on complete radiology reports, we propose a method to identify abnormal findings from the reports in addition to grouping them with unsupervised clustering and minimal rules. We formulate the task as cross-modal retrieval and propose Conditional Visual-Semantic Embeddings to align images and fine-grained abnormal findings in a joint embedding space. We demonstrate that our method is able to retrieve abnormal findings and outperforms existing generation models on both clinical correctness and text generation metrics.

1 Introduction

Understanding abnormal findings on radiographs (e.g. chest X-Rays) is a crucial task for radiologists. There has been growing interest in automatic radiology report generation to alleviate the workload of radiologists and improve patient care. Following the success of neural network models in image-to-text generation tasks (e.g. image captioning), researchers have trained CNN-RNN encoder-decoder networks to generate reports given radiology images (Shin et al., 2016; Kougia et al., 2019).

Although such models are able to generate fluent reports, the generation quality is often limited by biases introduced from training data or the training process. Figure 1 shows an example of chest X-rays (CXRs) and the associated reports from a public dataset (Johnson et al., 2019), along with the outputs generated by different models.[1] One issue is that models trained on complete reports tend to generate normal findings as they dominate the dataset (Harzig et al., 2019); another issue is that such generation models struggle to generate long and diverse reports as in other natural language generation (NLG) tasks (Boag et al., 2019).

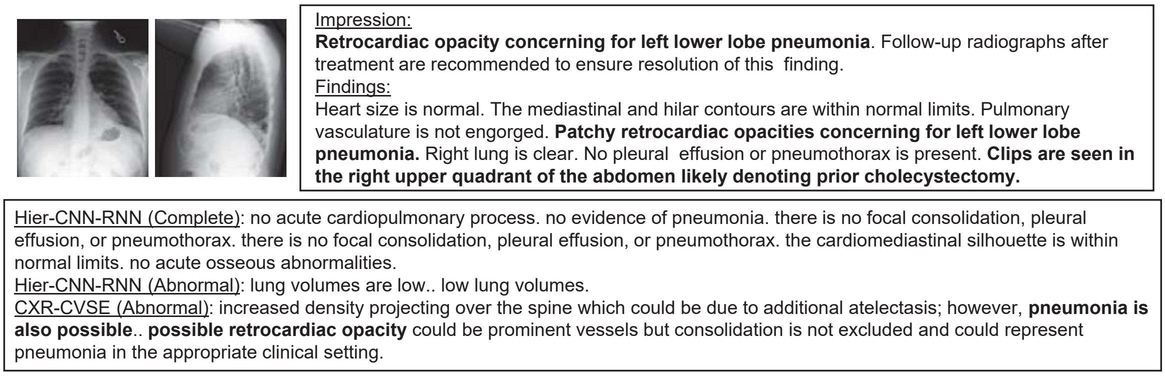
In this work, we focus on reporting abnormal findings on radiology images which are of higher importance to radiologists. To address issues of data bias, we propose a method to identify abnormal findings from existing reports and further use K-Means plus minimal mutual exclusivity rules to group these abnormal findings, which reduces the substantial burden of curating templates of abnormal findings. Given the fact that radiology reports are highly similar and have a limited vocabulary (Gabriel et al., 2018), we propose a cross-modal retrieval method to capture relevant abnormal findings from radiology images. Our contributions are summarized as:

- We learn conditional visual-semantic embeddings on radiology images and reports, which can be used to measure the similarity between image regions and abnormal findings by optimizing a triplet ranking loss.
- We develop an automatic approach to identify and group abnormal findings from large collections of radiology reports.
- We conduct comprehensive experiments to show that our retrieval-based method trained on the abnormal findings largely outperforms encoder-decoder generation models on clinical correctness and NLG metrics.

[*] Now at Google

[1] For a CXR report, 'Findings' is a detailed description and the 'Impression' is a summary.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1954–1960
November 16 - 20, 2020. ©2020 Association for Computational Linguistics



Impression:

Retrocardiac opacity concerning for left lower lobe pneumonia. Follow-up radiographs after treatment are recommended to ensure resolution of this finding.

Findings:

Heart size is normal. The mediastinal and hilar contours are within normal limits. Pulmonary vasculature is not engorged. **Patchy retrocardiac opacities concerning for left lower lobe pneumonia.** Right lung is clear. No pleural effusion or pneumothorax is present. **Clips are seen in the right upper quadrant of the abdomen likely denoting prior cholecystectomy.**

Hier-CNN-RNN (Complete): no acute cardiopulmonary process. no evidence of pneumonia. there is no focal consolidation, pleural effusion, or pneumothorax. there is no focal consolidation, pleural effusion, or pneumothorax. the cardiomediastinal silhouette is within normal limits. no acute osseous abnormalities.

Hier-CNN-RNN (Abnormal): lung volumes are low.. low lung volumes.

CXR-CVSE (Abnormal): increased density projecting over the spine which could be due to additional atelectasis; however, **pneumonia is also possible**.. **possible retrocardiac opacity** could be prominent vessels but consolidation is not excluded and could represent pneumonia in the appropriate clinical setting.

Figure 1: Example of CXR images (frontal and lateral views) and the associated report. Bolded are abnormal findings in the ground-truth and predictions. The CNN-RNN model trained on the complete reports tends to generate normal findings. Both CNN-RNN models generate repetitive sentences.

2 Related Work

2.1 Hierarchical encoder-decoder models

Jing et al. (2017) proposed a co-attention based Hierarchical CNN-RNN model that jointly trains two tasks: report generation and Medical Text Indexer (MTI) prediction. The model first predicts MTI tags and the semantic embeddings of the predictions are fed into the cascaded decoder for generation. Similarly, Yuan et al. (2019) extracted medical concepts from the CXR reports using SemRep[2] as alternatives to MTI tags. To address data bias, Harzig et al. (2019) proposed a CNN-RNN model with dual word-level decoders: one for abnormal findings and the other for normal findings. It jointly predicts whether the next sentence is a normal or abnormal finding, and uses the corresponding decoder to generate the next sentence. However, it still formulates the task as text generation and has the limitations of such models.

2.2 Hybrid retrieval-generation models

There has been increasing interest in studying hybrid retrieval-generation models to complement generation. Li et al. (2018) introduced a hybrid retrieval-generation framework which decides at each step whether it retrieves a template or generates a sentence. Li et al. (2019) proposed a model based on abnormality graphs, which first predicts existing abnormalities on the radiology images, then retrieves and paraphrases the templates of that abnormality. However, such models usually require non-trivial human effort to construct high quality prior knowledge (e.g. sentence templates, abnormality terms). Unlike previous work, we leverage unsupervised methods and minimal rules to group sentences into different abnormality clusters, seeking to minimize human effort.

2.3 Visual-semantic embeddings for cross-modal retrieval

Learning visually grounded semantics to facilitate cross-modal retrieval (i.e., image-to-text and text-to-image) is a challenging task for cross-modal learning (Faghri et al., 2018; Wu et al., 2019). Different from image captioning tasks, radiology reports are often longer and consist of multiple sentences, each related to different abnormal findings; meanwhile, there are fewer distinct objects in radiology images and the differences among images are more subtle.

3 Approach

Given radiology images I_f and I_l from the frontal and lateral view, Hierarchical CNN-RNN based methods predict complete medical reports $R = \{s_1, s_2, \ldots, s_N\}$, consisting of N sentences. Each sentence s_i is generated hierarchically:

$$P(s_i) = \prod_{t=1}^{T_i} P(w_i^t | w_i^{<t}, s_{<i}, E_f, E_l), \qquad (1)$$

where E_f and E_l are the feature maps of the images I_f and I_l generated by the CNN encoder, and w_i^t is the t-th word at the i-th sentence.

Instead of training such generation models, we approach the task as a cross-modal retrieval method. In particular, we propose a model that (1) measures the similarity between images and abnormal find-

[2] https://semrep.nlm.nih.gov/

ings, and (2) identifies fine-grained relevant image regions for each abnormal finding.

3.1 Problem definition

Assume each report $R_a = \{a_1, a_2, \ldots, a_M\}$ includes M abnormal findings (i.e., sentences). R_a is a subset of the complete report $R = \{s_1, s_2, \ldots, s_N\}$, where s_i can either be an abnormal sentence a_i or not.

Let $v \in \mathbb{R}^{d1}$ be the semantic embedding of an abnormal finding a of this report, and $E = \{m_j \in \mathbb{R}^{d2}\}_{j=1}^{w \times h}$ be the feature maps of the radiology image I associated with R_a, where j means the j-th region of the feature map. We first transform them into the joint embedding space $\mathbb{R}^d$ with separate linear projection layers:

$$\mathbf{v} = \text{norm}(\text{linear}(v)); \mathbf{m}_j = \text{norm}(\text{linear}(m_j)),$$

where we apply l_2 normalization on the joint embeddings to improve training stability, following work in visual-semantic embeddings (Faghri et al., 2018).

Next, we need to measure the similarity between the semantic and visual embeddings. As different regions may include details about different abnormal findings, we propose Conditional Visual-Semantic Embeddings (CVSE) to learn the fine-grained matching between regions and a target abnormal finding:

$$\begin{aligned} d(a, I) &= -\sum_{1 \le j \le w \times h} \alpha_j ||\mathbf{m}_j - \mathbf{v}||^2, \\ \hat{\alpha_j} &= \mathbf{v_\alpha}^\top (W_\alpha[\mathbf{m}_j; \mathbf{v}] + \mathbf{b}_\alpha), \\ \boldsymbol{\alpha} &= \text{softmax}(\hat{\boldsymbol{\alpha}}), \end{aligned} \tag{2}$$

where α_j is the attention score that represents the relevance between the region $\mathbf{m}_j$ and the abnormal finding $\mathbf{v}$, $d(a, I)$ is the similarity score between image I and the abnormal finding a, which is calculated as an attention-weighted sum over the similarity scores of each region with the abnormal finding. We use the (negative) squared l_2 distance to measure similarity. Since each report has both frontal and lateral views, the final similarity score is calculated as the average:

$$d_*(a, I) = \frac{1}{2}(d(a, I_f) + d(a, I_l)). \tag{3}$$

Finally, we optimize the hinge-based triplet ranking loss to learn the visual-semantic embeddings:

$$\begin{aligned} \mathcal{L} = &\sum_I [d_*(a^-, I) - d_*(a^+, I) + \delta]_+ \\ &+ \sum_a [d_*(a, I^-) - d_*(a, I^+) + \delta]_+, \end{aligned} \tag{4}$$

where δ is the margin, $[x]_+ = max(x, 0)$ is the hinge loss, a^+ (I^+) denotes a matched abnormal finding (image) from the training set while $a^-(I^-)$ denotes an unmatched abnormal finding (image) sampled during training.

3.2 Extracting and clustering abnormal findings

To identify abnormal findings in radiology reports, we train a sentence-level classifier which determines whether a sentence includes abnormal findings or not. We fine-tuned BERT (Devlin et al., 2019) on an annotated sentence-level dataset released by Harzig et al. (2019), which is a labeled subset of the Open-I dataset (Demner-Fushman et al., 2016). We achieve an F1-score of 98.3 on the held-out test set. We then use it to distantly label the reports from the MIMIC-CXR dataset (Johnson et al., 2019), which is the largest public CXR imaging report dataset.

Given that most medical reports are written following certain templates, many abnormal findings are often paraphrases of each other. We obtain the sentence embeddings via pre-trained models and apply K-Means to cluster the sentences about similar abnormal findings into 500 groups. We also design several simple mutual exclusivity rules to refine the groupings. We consider critical attributes such as position (e.g. left, right), severity (e.g. mild, severe) which often are not present at the same time. Then we apply these rules to separate each group formed by K-Means. Ultimately, we obtained 1,306 groups of abnormal findings.

4 Experiments

We compare CVSE with the state-of-the-art report generation models and simple baseline models to answer two research questions—**RQ1**: Does our retrieval-based method outperform generation models? **RQ2**: Do the visual-semantic embeddings capture abnormal findings grounded on images?

4.1 Baselines

We consider (1) the Hier-CNN-RNN model (Jing et al., 2017; Liu et al., 2019), as denoted in eq. (1);

Table 1: Comparisons of different models' clinical accuracy and NLG metrics. Accuracy, precision and recall are the macro-average across all 14 diseases.

Model	Accuracy	Precision	Recall	BLEU-4	BLEU-1	ROUGE-L	METEOR
MIMIC-CXR (Abnormal)							
CVSE + mutual exclusivity	**0.863**	**0.317**	**0.224**	**0.036**	0.192	**0.153**	0.077
CVSE	0.856	0.303	0.218	0.032	**0.197**	0.153	**0.088**
Hier-CNN-RNN	0.850	0.261	0.157	0.019	0.084	0.149	0.059
Hier-CNN-RNN + shuffle	0.853	0.172	0.117	0.013	0.064	0.130	0.046
MIMIC-CXR (Complete)							
Hier-CNN-RNN + complete	0.835	0.145	0.135	0.096	0.258	0.257	0.121
Hier-CNN-RNN + co-attention	0.843	0.156	0.127	0.098	0.281	0.252	0.120
Hier-CNN-RNN + dual	0.843	0.194	0.142	0.095	0.282	0.256	0.123

(2) Hier-CNN-RNN + co-attention (Jing et al., 2017) with co-attention on both the images and the predicted medical concepts; (3) Hier-CNN-RNN + dual, with the dual word-level decoders (Harzig et al., 2019). We also implement two simple variants: (4) Hier-CNN-RNN + complete, which considers the complete medical reports (i.e., both normal and abnormal findings) as input; (5) Hier-CNN-RNN + shuffle, whose input reports have a shuffled sentence order. Vinyals et al. (2015) has shown that input order affects the performance for encoder-decoder models and (5) could potentially address the training issue due to the static input order.

In all experiments, the abnormal set and complete set consist of the same (image, report) pairs. As discussed in Section 3.1, the abnormal set only considers the abnormal finding sentences of the report, which is a subset of sentences of the complete report. We compare these two sets to show that models trained on the abnormal sentences would achieve substantial improvement than those trained on the complete reports, which has not been studied before.

We use the CheXpert labeler to evaluate the clinical accuracy of the abnormal findings reported by each model, which is the state-of-the-art medical report labeling system (Irvin et al., 2019; Johnson et al., 2019). Given sentences of abnormal findings, CheXpert will give a positive and negative label for 14 diseases. We then calculate the Precision, Recall and Accuracy for each disease based on the labels obtained from each model's output and from the ground-truth reports.

4.2 Implementation details

We consider CXRs from the MIMIC-CXR dataset with both frontal and lateral views which include at least one abnormal finding. Ultimately, we obtain 26,946/3,801/7,804 CXRs for the train/dev/test sets, respectively. For the CVSE model, we set α to 0.2 and for each sample we randomly pick 8 negative samples. We use the pre-trained DenseNet-121 to obtain the feature maps of the CXR images. We use the pre-trained biomedical sentence embeddings (Zhang et al., 2019) to obtain initial embeddings for the abnormal findings.[3] The final dimension of the joint embedding d is set to 512. We take the top 3 retrieval results as the predicted abnormal findings. For all CNN-RNN based models, we use a VGG-19 model as the encoder, a 1-layer LSTM as the sentence decoder and a 2-layer LSTM as the word decoder. All dimensions are set to 512. Greedy search is applied during the decoding stage, following Jing et al. (2017). Our code are available online.[4]

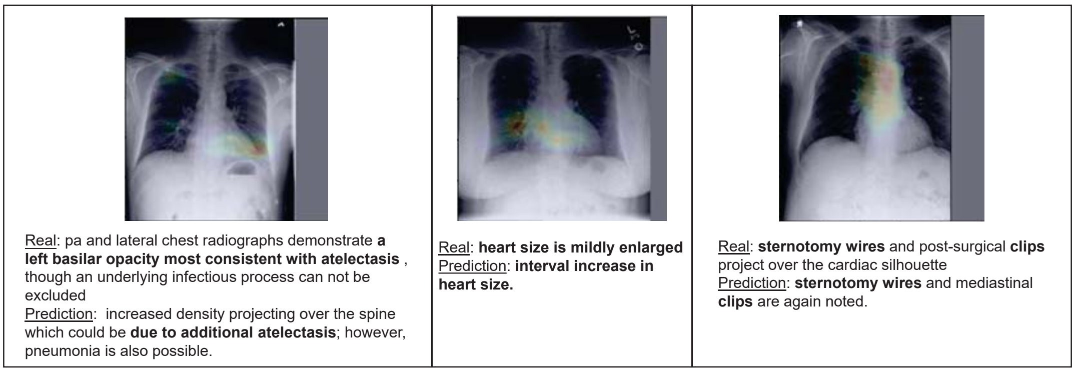
4.3 Performance comparison

We conduct experiments on both the abnormal and complete set of the MIMIC-CXR dataset which consider the abnormal findings in reports and the complete reports, respectively. As shown in Table 1, adding co-attention over medical concepts and dual decoders both improve the vanilla Hier-CNN-RNN model's clinical accuracy on the complete dataset. However, simply training the Hier-CNN-RNN model on the abnormal set would achieve better clinical accuracy. This shows the importance of addressing dataset bias. We also observe that the Hier-CNN-RNN model with a shuffled sentence order doesn't improve performance, which indicates the difficulty of addressing order bias during training of encoder-decoder models.

Our CVSE model outperforms all baselines on clinical accuracy metrics, which demonstrates its capability to accurately report abnormal findings. Notably, CVSE achieves significant improvements

[3] https://github.com/ncbi-nlp/BioSentVec

[4] https://github.com/nijianmo/chest-xray-cvse



Real: pa and lateral chest radiographs demonstrate **a left basilar opacity most consistent with atelectasis** , though an underlying infectious process can not be excluded
Prediction: increased density projecting over the spine which could be **due to additional atelectasis**; however, pneumonia is also possible.

Real: **heart size is mildly enlarged**
Prediction: **interval increase in heart size.**

Real: **sternotomy wires** and post-surgical **clips** project over the cardiac silhouette
Prediction: **sternotomy wires** and mediastinal **clips** are again noted.

Figure 2: Visualization of the attention maps from our method. 'Real' and 'Prediction' indicates the ground-truth and predicted abnormal findings.

on precision and recall. On the other hand, the baseline models will always miss abnormal findings thus leading to 0 precision and recall for many disease classes. More detailed results are included in the *appendices*.

Refining the groups with mutual exclusivity rules further improves the performance of CVSE. We also report the automatic evaluation of NLG metrics. As shown in Table 1, CVSE achieves higher scores than other baselines on the abnormal set.[5]

4.4 Qualitative analysis

We performed a human evaluation in which we sampled 20 images and asked a board-certified radiologist to give Likert scores (1 to 10) based on how closely the results generated by the model relate to the input images. The ground-truth obtained an average score of 7.85; our CVSE achieved a score of 6.35, higher than Hier-CNN-RNN trained on the abnormal set which obtained 6.15. The radiologist commented that Hier-CNN-RNN's outputs were simpler predictions, with less details; meanwhile, CVSE covered more abnormalities but may included false information sometimes.

In Figure 2, we visualize the attended regions on CXRs to investigate what part is important for reporting abnormal findings. We observe that our attention mechanism is able to detect relevant regions (e.g. heart, left opacity, wires) to determine which abnormal findings reside in the CXRs.

5 Conclusions

In this paper, we study how to build assistive medical imaging systems that report abnormal findings on the medical images in the form of detailed descriptions. We formulate the problem as a cross-modal retrieval task and apply a metric learning-based method to align visual and semantic features (i.e., image regions and textual descriptions of abnormal findings) without explicit labels. Our experiments show that the retrieval-based method outperforms generation-based models by mitigating their weaknesses in generating repetitive sentences and bias toward normal findings. In the future, we will extend our method to other medical image datasets and explore transfer learning.

Acknowledgments

This work was supported in part by the Office of the Assistant Secretary of Defense for Health Affairs through the Accelerating Innovation in Military Medicine Research Award program under Award No. W81XWH-20-1-0693, and NSF #1750063.

References

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alsentzer, and Peter Szolovits. 2019. Baselines for chest x-ray report generation. In *ML4H*.

Dina Demner-Fushman, M. Kohli, M. Rosenman, S. E. Shooshan, Laritza Rodriguez, S. Antani, G. Thoma, and C. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2:304–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-

[5] Models trained on the complete set can match the predominant normal findings thus leading to higher NLG metrics.

semantic embeddings with hard negatives. In *BMVC*.

Rodney A. Gabriel, Tsung-Ting Kuo, Julian J. McAuley, and Chun-Nan Hsu. 2018. Identifying and characterizing highly similar notes in big clinical note datasets. *Journal of biomedical informatics*, 82:63–69.

Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. 2019. Addressing data bias problems for chest x-ray image report generation. In *BMVC*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2017. On the automatic generation of medical imaging reports. In *ACL*.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr: A large publicly available database of labeled chest radiographs. *ArXiv*, abs/1901.07042.

Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. *ArXiv*, abs/1905.13302.

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

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *ArXiv*, abs/1903.10122.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *MLHC*.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2015. Order matters: Sequence to sequence for sets. In *ICLR*.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *CVPR*.

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

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6.

A Implementation details

A.1 Mutual exclusive rules to refine groupings

Though advanced sentence embedding methods allow for effective groupings of sentences in radiology reports describing similar clinical features, they fail to distinguish antonyms such as right vs. left because antonyms share highly similar contexts and are considered to be semantically similar by these embedding methods. For our purposes, however, it is important to distinguish some of the antonyms because they describe mutually exclusive image features. For example our grouping based on a sentence embedding results clustered these sentences in the same group:

- *continued right lung volume loss.*
- *there is right lung volume loss again noted.*
- *right lung volume loss is again noted.*
- *there is volume loss of the left upper lung.*
- *left upper lobectomy changes including left lung volume loss.*
- *left upper lobe volume loss is present.*

To separate those denoting right lung volume loss from those denoting left we wrote simple matching rules to identify selected words in sentences in the same group that are mutually exclusive and encode their occurrences as one-hot vectors. Then we applied the DBSCAN clustering method in the sklearn[6] library to divide the group further into on average three subgroups based on the one-hot vector encoding. We considered six sets of mutually exclusive terms:

- right, left, bilateral.
- small, great|large.
- low, high.
- elevate|enlarge|increase|widen, shrink|decrease.

[6] https://scikit-learn.org/stable/

Table 2: Detailed Accuracy, precision and recall for different models.

Model	CVSE + mutual exclusiveness			Hier-CNN-RNN (abnormal)		
Disease	Accuracy	Precision	Recall	Accuracy	Precision	Recall
No Finding	**0.769**	**0.346**	**0.265**	0.766	0.336	0.259
Enlarged Cardiomediastinum	0.926	**0.063**	**0.060**	**0.959**	0.000	0.000
Cardiomegaly	0.801	0.512	**0.606**	**0.813**	**0.570**	0.338
Lung Lesion	0.921	**0.192**	**0.121**	**0.943**	0.000	0.000
Lung Opacity	**0.692**	**0.635**	**0.237**	0.658	0.500	0.021
Edema	0.920	0.405	**0.206**	**0.927**	**0.490**	0.084
Consolidation	0.876	**0.130**	**0.181**	**0.935**	0.079	0.006
Pneumonia	**0.859**	**0.364**	**0.214**	0.855	0.306	0.154
Atelectasis	**0.773**	**0.525**	0.320	0.599	0.284	**0.469**
Pneumothorax	0.964	**0.073**	**0.051**	**0.977**	0.000	0.000
Pleural Effusion	**0.894**	**0.640**	0.465	0.696	0.262	**0.703**
Pleural Other	0.962	**0.145**	**0.036**	**0.968**	0.000	0.000
Fracture	0.917	0.063	**0.050**	**0.935**	**0.072**	0.029
Support Devices	0.808	0.348	**0.321**	**0.863**	**0.752**	0.130
Macro-Average	**0.863**	**0.317**	**0.224**	0.850	0.261	0.157

- improve|resolve|clear, worsen.
- mild, severe.

A.2 Parameter settings

We use PyTorch to implement all models and run them on 2 1080Ti GPUs. We resize all images into size of 512×512 for both models. For all experiments, we save the models that perform best on the validation set. For CVSE, we measure recall on validation set; for CNN-RNN models, we consider perplexity on validation set.

For CVSE we use an Adam optimizer with a learning rate 0.001 and training continues for 40 epochs. For all Hier-CNN-RNN models, we set the learning rate for encoder and decoder as $5e^{-6}$ and $2e^{-4}$, respectively. We train the models for 100 epochs. We use a VGG-19 model as the encoder, a 1-layer LSTM as the sentence decoder and a 2-layer LSTM as the word decoder. We observe slightly better performance from VGG-19 compared to DenseNet-121 for the generation models. For models that require medical concepts, we use SemRep (i.e. a UMLS-based program released by NIH) to extract 93 highly frequent medical concepts from the training set.

B Experiments on MIMIC-CXR

B.1 Detailed clinically accuracy results on 14 diseases

Table 2 shows the detailed accuracy, precision and recall on all 14 diseases from our CVSE model with mutual exclusiveness rules and the Hier-CNN-RNN model trained on the abnormal set. Overall, CVSE outperforms Hier-CNN-RNN on the macro-average of accuracy, precision and recall. Notably, CVSE achieves higher recall on 12 out of 14 diseases with a comparative or higher precision. Meanwhile, Hier-CNN-RNN outputs 0 positive predictions on 4 disease types that are dominated by the negative findings, which shows its limited capability to generate diverse predictions.