Baselines for Chest X-Ray Report Generation

William Boag WBOAG@MIT.EDU
MIT CSAIL

Tzu-Ming Harry Hsu STMHARRY@MIT.EDU
MIT CSAIL

Matthew McDermott MMD@MIT.EDU
MIT CSAIL

Gabriela Berner GBERNER@COLLEGE.HARVARD.EDU
Harvard

Emily Alesentzer EMILYA@MIT.EDU
MIT CSAIL

Peter Szolovits PSZ@MIT.EDU
MIT CSAIL

Abstract

With advances in deep learning and image captioning over the past few years, researchers have recently begun applying computer vision methods to radiology report generation. Typically, these generated reports have been evaluated using general domain natural language generation (NLG) metrics like CIDEr and BLEU. However, there is little work assessing how appropriate these metrics are for healthcare, where correctness is critically important. In this work, we profile a number of models for automatic report generation on this dataset, including: random report retrieval, nearest neighbor report retrieval, n-gram language models, and neural network approaches. These models serve to calibrate our understanding for what the opaque general domain NLG metrics mean. In particular, we find that the standard NLG metrics (e.g. BLEU, CIDEr) actually assign higher scores to random (but grammatical) clinical sentences over n-gram-derived sentences, despite the n-gram sentences achieving higher clinical accuracy. This casts doubt on the usefulness of these domain-agnostic metrics, though unsurprisingly we find that the best performance – on both CIDEr/BLEU and clinical correctness – was achieved by more sophisticated models.

1. Introduction

The automatic processing of radiological images and their associated free-text reports (e.g., Figure 1) is one of the most rapidly growing areas of Machine Learning for Healthcare. This is in part motivated by the tremendous success ML has shown in other areas of computer vision, and in part by the large potential for ML-aided assistive technologies within the clinical radiological workflow. The stream of recent FDA approvals of machine learning algorithms for radiology applications, including x-ray wrist fracture diagnosis and brain MRI interpretation, is a testament to the increasing desire to bring these methods to the clinic Topol (2019).

The automatic generation of written reports from radiology images has received particular attention, as generation of reports allows for interpretable predictions of images that can be easier than multi-label classifications for clinicians to understand. Additionally, natural language more naturally fits into existing clinician workflows. Radiology report generation has been explored in a number of ways, with models ranging in scale and complexity. However, these models are largely performed on closed datasets or partially closed datasets (i.e. not containing publicly-available notes). While there have been several previously released public radiology datasets Demner-Fushman et al. (2016); Wang et al. (2017); Irvin et al. (2019); Bustos et al. (2019), these datasets — with one very recent exception Bustos et al. (2019) — do not contain public radiology reports, making it difficult for researchers to reproduce published results and truly understand best practices in terms of model architectures and principles. And of that exception, those reports are not in English.

The new release of the MIMIC-CXR dataset Johnson et al. (2019) – the first major U.S. release of a dataset containing paired images and free-text reports – changes this. This dataset aims to enable researchers to truly iterate on each-others work and more rapidly advance the state-of-the-art for automatic radiology report generation, just as was done in the general domain. In order to understand the nature and challenges of this task, it is essential to have strong baselines Boag (2019). In this work, we employ several baseline methods, including n-gram language models, nearest neighbor models, and neural network approaches for automatic report generation. We profile these methods quantitatively both with standard text generation evaluation metrics (BLEU and CIDEr) and with the CheXpert clinical report labeling system to asses their clinical accuracy. Our code and experiments are publicly available for anyone with access to MIMIC-CXR[1].

We find that neural network approaches offer the strongest performance on these models, but 1-NN methods are robust contenders, specifically with regards to clinical efficacy as measured by CheXpert predicted label F1 scores. In light of this – and the ease with which one can implement nearest neighbor methods – we recommend including a 1-NN baseline for future report generation projects as a new best practice. This will help disentangle the benefits of better image encoding vs better feature-to-text decoding. This is especially relevant because we expect that with transformer-based language modeling, it is likely that feature-to-text decoding will improve in the near future. Additionally, our qualitative eval-

[1] https://github.com/wboag/cxr-baselines

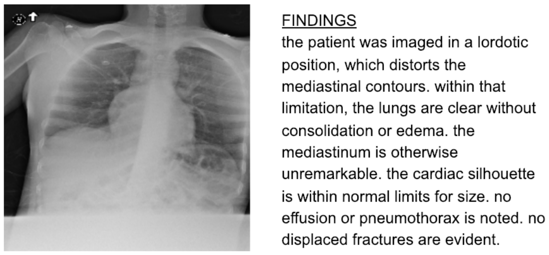
FINDINGS
the patient was imaged in a lordotic position, which distorts the mediastinal contours. within that limitation, the lungs are clear without consolidation or edema. the mediastinum is otherwise unremarkable. the cardiac silhouette is within normal limits for size. no effusion or pneumothorax is noted. no displaced fractures are evident.

Figure 1: A radiological image (Chest X-Ray) and the associated natural language description of the findings, written by a radiologist.

Dataset	Description	# Radiographs	# Reports	# Patients
Demner-Fushman et al. (2016) Open-I	Open-I is a modest dataset of chest x-ray images and reports from the Indiana Network for Patient Care.	8121	3996	3996
Wang et al. (2017) NIH Chest-XRay8	NIH Chest-XRay8 contains clinically labeled chest radiographs. The labels were determined algorithmically, not via clinician annotation.	108,948	0	32,717
Irvin et al. (2019) CheXpert	CheXpert, like NIH Chest-XRay8, contains algorithmically labled chest radiographs.	224,316	0	65,240
Bustos et al. (2019) PadChest	PadChest is a large Spanish dataset containing chest radiographs, free-text reports, and highly granular, algorithmically determined labels.	160,868	109,931	67,625
Johnson et al. (2019) MIMIC-CXR	MIMIC-CXR is the largest public dataset containing both chest radiographs and free-text reports. Clinical labels produced via CheXpert, can also be used.	473,057	206,563	63,478

Table 1: A description of available chest x-ray datasets.

uation reveals that while neural network approaches tend to generate aggressively short, simple, and "normal" sentences, n-gram approaches instead, generate longer but surprisingly readable sentences. Lastly, we find that general domain NLG metrics (e.g. BLEU, CIDEr) actually assign higher scores to random (but grammatical) clinical sentences over n-gram-derived sentences, despite the n-gram sentences achieving higher clinical accuracy.

2. Related Works

Chest Radiograph Datasets In recent years, several chest radiograph datasets totalling over half a million x-ray images have been made publicly available. A summary of these datasets is available in Table 1. The notes for PadChest are multilingual.

Radiology Images & Reports in General Researchers have explored using ML techniques on radiographs and their associated free-text reports in a number of ways. Working from images alone, researchers have successfully identified common thorax diseases via chest

X-Rays Rubin et al. (2018), detecting metastatic breast cancer Wang et al. (2016), classifying hip fractures Gale et al. (2017), and detecting pneumonia Rajpurkar et al. (2017).

Lastly, researchers have explored using these two modalities in concert, which is expected to help further improve the model performance in both image annotation and automatic report generation Litjens et al. (2017). Researchers have examined joint embedding spaces for information retrieval Hsu et al. (2018), multimodal processing for improved tissue classification, both over optical coherence tomography images Schlegl et al. (2015) as well as for annotation of anatomy, disease state, and severity for chest X-rays Shin et al. (2016). More recently, researchers have examined using multimodal information to produce more interpretable models via saliency maps over chest X-ray images Moradi et al. (2018). Additionally, these approaches have been explored for simultaneous classification and report generation Wang et al. (2018).

Image Captioning Over the last six years, Computer Vision (CV) has undergone a revolution. With the success of AlexNet Krizhevsky et al. (2012), researchers began using deep learning architectures to massively improve performance. After achieving sufficient performance on image classification, the community began exploring other tasks, including the intersection of CV and Natural Language Processing (NLP). In 2014, Microsoft released its COCO Lin et al. (2014) dataset for image captioning, which received significant attention and effort. This dataset contained millions of images, each annotated with at least five human-written captions describing what is happening. The task is to use the image as input to generate a readable, accurate, and linguistically correct caption.

Works such as *Show and Tell* Vinyals et al. (2015) received tremendous acclaim for their success at this task. Once a leaderboard was established, other works also proved effective, including boosting methods Yao et al. (2016) and policy gradient optimization Rennie et al. (2017). Interestingly, one of the top-performing systems was a nearest neighbor approach Devlin et al. (2015) which recognized the redundancy of simple labels in the dataset such as "A car parked in front of a building." Relatedly, the *Show and Tell* paper found that up to 80% of the their approach's generated captions were identical to training dataset captions in MS COCO, yet they still achieved near state-of-the-art performance.

Medical Text Generation There is a long history of research to generate notes from electronic health record (EHR) data Pivovarov and Elhadad (2015). This task is complex because it involves both identifying relevant information from noisy, heterogeneous EHR data and then synthesizing the information into human-readable text. While a number of approaches have relied on rules based systems and structured domain knowledge to generate nursing shift summaries Hunter et al. (2012) and discharge summaries Goldstein and Shahar (2015), more recent work has tackled this problem using an end-to-end deep learning system that generates notes directly from demographics, previous notes, labs, and medications Liu (2018).

For report generation, Jing et al. (2017) built a multi-task learning framework, which includes a co-attention mechanism module, and a hierarchical long short term memory (LSTM) module, for radiological image annotation and report paragraph generation. Li et al. (2018a) proposed a reinforcement learning-based Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) to learn a report generator that can decide whether to retrieve a template or generate a new sentence. Alternatively, Gale et al. (2018) generated the interpretable hip fracture X-ray reports by identifying image features and text templates filling. Finally, Liu et al. (2019) generates reports for chest X-rays using a hierarchical neural approach, augmented with a reinforcement learning penalty to maintain clinical accuracy.

3. Methods

3.1. Data

All experiments in this work used the MIMIC-CXR dataset. MIMIC-CXR consists of 473,057 chest x-ray images and 206,563 reports from 63,478 patients. Of these images, 240,780 are of anteroposterior (AP) views, which we focus on in this work. Further, we eliminate all duplicated radiograph images with adjusted brightness or contrast[2], leaving a total of 95,242/87,353 images/reports, which we subdivide into a train set of 75,147/69,171 and a test set of 19,825/18,182 images/reports, with no overlap of patients between the two. Radiological reports are parsed into sections and we use the findings section. See Figure 1 for an example chest x-ray and report.

3.2. Models

We experiment with models of a range of complexity. For all image-conditional models, we represent images according to the induced representation by a deep convolutional neural network (CNN). Per CheXNetRajpurkar et al. (2017), we use a DenseNet121 to extract features of size $8 \times 8 \times 1024$, which are then globally mean-pooled to a final, 1024-dimensional representation. The networks were pretrained using ChestX-ray14 classification tasks. Wang et al. (2017)

We tested a variety of language generation models.

Random Retrieval Baseline This is our simplest tested baseline. It is unconditioned upon the query image, and instead merely draws a random report from the training set. This selected report is treated as the "generated" text. These reports will be readable, but not relevant to the query image at all, and are thus unlikely to score well either on the text-generation outputs (which compare overlapping n-grams to the true report), or on our measures of clinical relevance.

Conditional n-gram Language Model N-gram language models rely on the Markov assumption that the next word in a sentence depends only on what the previous (N-1) words were and nothing else earlier. These models are learned from simply tallying how often a given word actually follows a given phrase in the training set. We make these models conditional on a query image by learning a per-instance language model for each image based on the reports corresponding to the closest 100 train images (in the DenseNet induced space).

Nearest Neighbor For this baseline, we "generate" our text by returning the caption of the training image with the largest cosine similarity (in the DenseNet-induced space) to the test query image. Like our random retrieval baseline, the generated captions are guaranteed to be grammatical. Additionally, here they should also be clinical relevant, which should increases both the standard NLG metrics as well as clinical accuracy measurements.

[2] Commonly produced for clinical needs

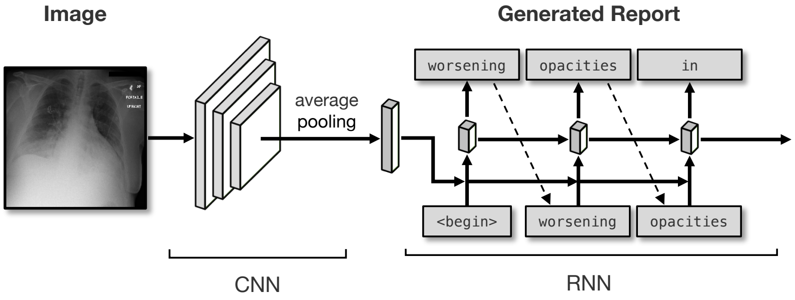

Figure 2: Full CNN-RNN pipeline.

CNN-RNN This is the canonical neural image captioning model pioneered in *Show and Tell* Vinyals et al. (2015). We employ a simple CNN encoder of the image, whose output is projected from 1024 dimensions to 256 dimensions to regularize the model, then fed into an LSTM decoder to produce the caption. The LSTM is trained to minimize the cross-entropy loss per token in the task of predicting the next word in the sentence. All weights, both the CNN's and the LSTM's were tuned via backpropagation. This model is shown in Figure 2.

The model is trained for 64 epochs with an initial learning rate of 1×10^{-3}, and the learning rate is decayed by 0.5 every 16 epochs. We train the model using *teacher forcing* Williams and Zipser (1989), which means that the model is trained initially feeding in only the true tokens into the decoder, rather than feeding in the decoded tokens. However, we increase the probability of feeding a sample of the inferred probability to itself by 0.05 per 16 epochs.

We employ a beam search decoding mechanism in addition to a greedy decoding strategy. A greedy decoder will simply always use the next token of highest probability given the generated history. However, this is sub-optimal, as it sacrifices the overall likelihood of a sentence (and thus the overall readability) for short term gains. A beam-search decoder, on the other hand, will always track the subsequent top-K most likely next tokens at each step, exploring a broader set of possible sentences. The final candidates are reweighted by a scoring function which encourages longer, more complex sentences, before finally returning and output sequence. We used a beam-search decoder with $K = 4$.

3.3. Evaluation

In this work, we report the standard BLEU and CIDEr scores to compare against existing approaches.

Some papers (TieNet) that generate captions also report accuracy of models by switching the LSTM generator with a multi-label classifier. This does assess how effective the CNN is at extracting the relevant features for diagnosis, but it does not actually measure whether the generated reports are correct. We assess the correctness of the generated reports by feeding them into the CheXpert sentence labeler to compare how often the generated reports agree with the clinical findings of the reference reports. Although the labeler does require a baseline level of grammaticality, this evaluation largely ignores readability or marginal increases in grammaticality. We assess correctness using CheXpert via accuracy, precision, and F1. F1 is a balanced measure of both precision and recall, and will allow us to better capture true performance here in light of the strong class imbalance of these datasets. CheXpert labels 14 categories of diseases and support devices. Metrics are calculated per category across all test set examples, and then the metrics are averaged to obtain a macro-average.

4. Results

Quantitative Results Table 2 shows the results of all models across the linguistic and clinical accuracy evaluation measures. The neural model with beam search (*CNN-RNN + Beam*) achieves the highest score across all measures of linguistic performance as well as CheXpert Accuracy and CheXpert Precision. However, it is defeated by a large margin (0.07 – greater than 25%) by the *1-NN* model in CheXpert F1. This indicates that the performance gain of the neural model disproportionately favors precision over recall (i.e. when it makes a claim, that claim is more likely true, at the expense of being overly cautious). Overall, *1-NN* performs relatively well at the linguistic measure, being competitive with (admittedly) basic neural approaches on a number of metrics.

We can more closely examine the clinical accuracy measures of these model outputs by examining the per-class F1 of each CheXpert category (Table 3). Here, the *1-NN* model achieves the highest F1 score for almost every single class. Unsurprisingly each method does better on higher-frequency labels, and this trend is consistent enough across models that the Macro-average and Micro-average F1s have the same ranking of models.

Perhaps unexpectedly, the *Random* approach also demonstrates non-trivial performance on these linguistic measures, outperforming both *3-gram* and *1-NN* on BLEU as well as *3-gram* on CIDEr as well. This should give some pause about what "good" performance on these tasks looks like: if models are not performing better than sampling irrelevant reports, then either the metric is bad or the models are not leaning anything.

One might suspect that perhaps BLEU and CIDEr are actually fine, and perhaps *3-gram* is simply generating nonsense. However we find that *3-gram* essentially ties the *CNN-RNN + Beam* on CheXpert F1 and surpasses *Random.* Despite achieving CNN-RNN-level clinical performance, the *3-gram* model is ranked lower than irrelevant-but-real reports by CIDEr and BLEU. This is likely because those metrics amount to overlapping ngram coverage, which favors surface-level grammatically, rather than correctness.

Table 2: Automatic evaluation metrics of baseline methods for image captioning task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	CheXpert Accuracy	CheXpert Precision	CheXpert F1
Random	0.265	0.137	0.070	0.036	0.570	0.770	0.146	0.148
1-gram	0.196	< 0.001	< 0.001	< 0.001	0.348	0.742	0.206	0.174
2-gram	0.194	0.098	0.043	0.013	0.404	0.764	0.225	0.193
3-gram	0.206	0.107	0.057	0.031	0.435	0.782	0.225	0.185
1-NN	**0.305**	0.171	0.098	0.057	0.755	0.818	0.253	**0.258**
CNN-RNN	0.004	< 0.001	< 0.001	< 0.001	0.066	0.822	0.144	0.067
CNN-RNN + Beam	**0.305**	**0.201**	**0.137**	**0.092**	**0.850**	**0.837**	**0.304**	0.186

Table 3: Per-class F1 of CheXpert sentence labels from generated reports.

	n	Random	3-gram	1-NN	CNN-RNN	CNN-RNN
Label						+ Beam
Support Devices	22227	0.316	0.388	0.527	0.106	**0.613**
Airspace Opacity	21972	0.038	0.326	**0.417**	0.330	0.077
Cardiomegaly	19065	0.113	0.390	**0.445**	0.022	0.390
Atelectasis	16161	0.241	0.271	**0.375**	0.054	0.146
No Finding	15677	**0.568**	0.286	0.455	0.362	0.407
Pleural Effusion	15283	0.222	0.364	**0.532**	< 0.001	0.473
Edema	6594	0.111	0.192	**0.286**	0.009	0.271
Enlarged Cardiomediastinum	6064	**0.234**	0.135	0.142	< 0.001	0.134
Pneumonia	3068	0.063	0.036	**0.080**	0.010	0.030
Pneumothorax	2636	0.043	0.082	**0.111**	0.042	0.043
Fracture	2617	0.041	0.022	**0.060**	< 0.001	< 0.001
Lung Lesion	2447	**0.277**	0.062	0.062	0.005	0.001
Consolidation	2384	0.043	0.037	**0.085**	0.002	0.014
Pleural Other	1285	0.020	< 0.001	**0.039**	< 0.001	< 0.001
Micro-Average	—	0.215	0.294	**0.397**	0.122	0.302
Macro-Average	—	0.148	0.185	**0.258**	0.067	0.186

The neural model absent beam search performs very poorly, demonstrating the importance of this more advanced decoding model. As indicated by both 2 and Table 3, beam search consistently adds a significant improvement across the board.

Qualitative Results We show qualitative results of each method on a few randomly chosen images in Figure 3. For the first image, we see that the *CNN-RNN + Beam* generates a very similar report as the reference, though the *3-gram* is also accurate (albeit verbose). We see consistently that the *1-NN* approach returns relevant captions which can be incorrectly over-specific about something that doesn't appear in the test image (e.g. the placement of a left pleural tube). This suggests potential limits in a neighbor-based approach for text generation when the target text is very unique (as opposed to ”a man is playing a guitar”).

5. Discussion

Despite the *CNN-RNN + Beam* outperforming the *3-gram* model across the board for BLEU and CIDEr, they have incredibly similar performance on the CheXpert task (0.185 vs 0.186 macro-average, as well as very similar performances as well for many classes). Though this is a limited study, this suggests that the simple neural model may not be any better at "describing" latent knowledge, and the limiting factor is the quality of the CNN-extracted features. Likely, it's a little of column-A and a little of column-B: neural models have been shown to benefit from additional decoding techniques (e.g. attention and hierarchical decoders), but it may also be worth developing stronger feature extractors for images. These extractors could then be tested in (closer to) isolation using simple decoding models like nearest neighbor and n-grams.

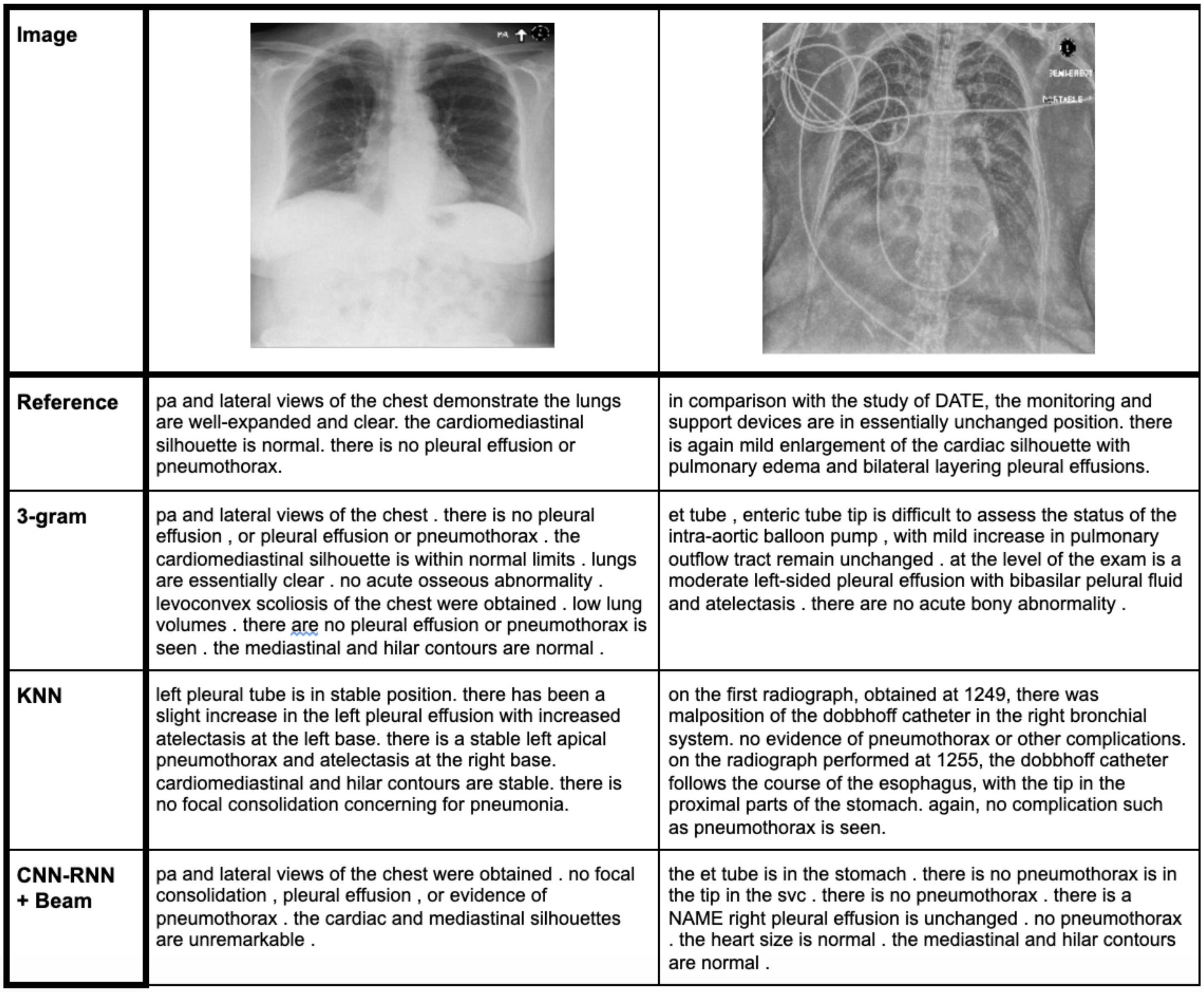

Image		
Reference	pa and lateral views of the chest demonstrate the lungs are well-expanded and clear. the cardiomediastinal silhouette is normal. there is no pleural effusion or pneumothorax.	in comparison with the study of DATE, the monitoring and support devices are in essentially unchanged position. there is again mild enlargement of the cardiac silhouette with pulmonary edema and bilateral layering pleural effusions.
3-gram	pa and lateral views of the chest . there is no pleural effusion , or pleural effusion or pneumothorax . the cardiomediastinal silhouette is within normal limits . lungs are essentially clear . no acute osseous abnormality . levoconvex scoliosis of the chest were obtained . low lung volumes . there are no pleural effusion or pneumothorax is seen . the mediastinal and hilar contours are normal .	et tube , enteric tube tip is difficult to assess the status of the intra-aortic balloon pump , with mild increase in pulmonary outflow tract remain unchanged . at the level of the exam is a moderate left-sided pleural effusion with bibasilar pelural fluid and atelectasis . there are no acute bony abnormality .
KNN	left pleural tube is in stable position. there has been a slight increase in the left pleural effusion with increased atelectasis at the left base. there is a stable left apical pneumothorax and atelectasis at the right base. cardiomediastinal and hilar contours are stable. there is no focal consolidation concerning for pneumonia.	on the first radiograph, obtained at 1249, there was malposition of the dobbhoff catheter in the right bronchial system. no evidence of pneumothorax or other complications. on the radiograph performed at 1255, the dobbhoff catheter follows the course of the esophagus, with the tip in the proximal parts of the stomach. again, no complication such as pneumothorax is seen.
CNN-RNN + Beam	pa and lateral views of the chest were obtained . no focal consolidation , pleural effusion , or evidence of pneumothorax . the cardiac and mediastinal silhouettes are unremarkable .	the et tube is in the stomach . there is no pneumothorax is in the tip in the svc . there is no pneumothorax . there is a NAME right pleural effusion is unchanged . no pneumothorax . the heart size is normal . the mediastinal and hilar contours are normal .

Figure 3: Example outputs from the best versions of each model type.

Additionally, these results point to limitations in text generation evaluation metrics. In the context of radiology report generation especially, we care simultaneously about both human readability *and* clinical accuracy. The ultimate test for this would be a prospective study to see what model produces the best care – for instance, it could well be the case that clinical correctness is most important so long as the report clears a threshold of being "good enough" to be readable – but of course any deployment would first require relatively high confidence of model correctness, which itself requires a proxy metric. However, our metrics at present are limited to either readability or correctness. Using both together as we do here is one solution, but there is an opportunity in this to design more holistic measures capable of capturing our true intentions.

Though these metrics are very commonly used, they have incurred significant push back as well because they tend to favor superficial, short sentencesBoag et al. (2016); Kilickaya et al. (2017). These concerns are likely especially true in the clinical domain, where we care not only about free-text readability, but also about the accuracy of the stated clinical conclusions. Further, these metrics were designed to be all-purpose tools, independent of any domain, which limits how reliable they might be expected to be for a highly specialized area such as medicine. It may prove true that these tools are sufficient proxies for even doctor judgment, but that has not yet been shown – these metrics were validated based on correlation with human judgment on generic sentences with a large number of reference sentences.

Limitations & Future Work This study has several notable limitations, each of which presents an opportunity for future work. First and foremost, many of the models we test are relatively simple. Even within our neural model, more advanced framing elements, such as attention or reinforcement learning – both of which have recently been used Wang et al. (2018); Li et al. (2018b) – could be explored here. Similarly, among our non-neural baselines, more aggressive measures of image-conditioning (e.g. use k-NN and take the centroid report to reduce over-specificity) and more complex language models could be employed. Second, CheXpert performance, while offering a valuable assessment of clinical performance, only assess a limited number of clinical categories and so doesn't capture the full battery of clinical concerns. Additionally, it relies on a rule-based parser for reports. Finally, in medical contexts we can often assess the dangers of various types of errors, as a type-1 vs. type-2 error can have very different consequences when diagnoses are concerned. Rather than merely relying on F1 to offer a more class-imbalance sensitive evaluation metric, we could instead attempt to analyze which kinds of errors are most critical here and use performance measures more sensitive to those relationships.

6. Conclusion

This study profiles a number of text-generation models for automatic radiology report generation across evaluation metrics spanning linguistic quality (BLEU-1 through BLEU-4, CIDEr) as well as clinical efficacy (CheXpert accuracy, precision, and F1).

These results demonstrate several important findings relating to automatic radiological report generation.

1. Beam search is critical for strong text realism in our neural model.

2. Though neural approach can excel at generating realistic text relative to traditional baselines, it may do so at the expense of clinical sensitivity. Here, our neighbor-based baseline consistently shows improved F1 performance across nearly every CheXpert class as compared to the neural model.

3. With a good image feature-extractor, even n-grams offer better performance than one might expect. The trigram model's Chexpert F1 is nearly identical to CNN-RNN + Beam.

4. Random retrieved reports score surprisingly highly at both linguistic and clinical evaluation metrics. This should help calibrate our understanding of how "good" it is for a model to achieve a certain score. This is very much an extension of the age-old wisdom that accuracy is a poor evaluation metric in cases of class imbalance (because a trivial predictor can still score high).

5. Standard NLG metrics are ill-equipped to measure the quality of clinical text. The over-relaince on n-gram overlap caused these metrics to favor irrelevant-but-fluent reports over correct-but-ungrammatical ones.

7. Acknowledgments

This research was funded in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374, NIH National Institutes of Mental Health grant P50-MH106933, a Mitacs Globalink Research Award, and Harvard Medical School Biomedical Informatics and Data Science Research Training Grant T15LM007092 (Co-PIs: Alexa T. McCray, PhD and Nils Gehlenborg, PhD).

References

William Boag. The baseline manifesto (blog). https://willieboag.wordpress.com/2019/10/21/the-baseline-manifesto. 2019.

William Boag, Renan Campos, Kate Saenko, and Anna Rumshisky. Mutt: Metric unit testing for language generation tasks. In *ACL*, August 2016.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv:1901.07441 [cs, eess]*, January 2019. URL http://arxiv.org/abs/1901.07441. arXiv: 1901.07441.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 23(2):304–310, March 2016. ISSN 1527-974X. doi: 10.1093/jamia/ocv080.

Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *CoRR*, abs/1505.04467, 2015. URL http://arxiv.org/abs/1505.04467.

William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P. Bradley, and Lyle J. Palmer. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv:1711.06504 [cs, stat]*, November 2017. URL http://arxiv.org/abs/1711.06504. arXiv: 1711.06504.

William Gale, Luke Oakden-Rayner, Gustavo Carneiro, Andrew P Bradley, and Lyle J Palmer. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*, 2018.

Ayelet Goldstein and Yuval Shahar. Generation of Natural-Language Textual Summaries from Longitudinal Clinical Records. *Studies in Health Technology and Informatics*, 216: 594–598, 2015. ISSN 9781614995630. doi: 10.3233/978-1-61499-564-7-594.

Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence In Medicine*, 56:157–172, 2012. doi: 10.1016/j.artmed.2012.09.002. URL https://ac-els-cdn-com.proxy.library.vanderbilt.edu/S0933365712001170/1-s2.0-S0933365712001170-main.pdf?_tid=fe240fec-175d-11e8-b627-00000aab0f26&acdnat=1519255490_b7e50dbe1b82becb5eb4307abefd9ac6.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie 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. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031 [cs, eess]*, January 2019. URL http://arxiv.org/abs/1901.07031. arXiv: 1901.07031.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

Alistair E. W. Johnson, Tom J. Pollard, Seth Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]*, January 2019. URL http://arxiv.org/abs/1901.07042. arXiv: 1901.07042.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. URL https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *arXiv preprint arXiv:1805.08298*, 2018a.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1530–1540. Curran Associates, Inc., 2018b. URL http://papers.nips.cc/paper/7426-hybrid-retrieval-generation-reinforced-agent-for-medical-image-report-generation.pdf.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014. URL https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

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. Ann Arbor, MI, August 2019. URL https://www.mlforhc.org/s/Liu_G.pdf. arXiv: 1904.02633.

P. J. Liu. Learning to write notes in electronic health records. *arXiv preprint arXiv:1808.02622*, 2018.

Mehdi Moradi, Ali Madani, Yaniv Gur, Yufan Guo, and Tanveer Syeda-Mahmood. Bimodal network architectures for automatic generation of image annotation from text. September 2018. URL http://arxiv.org/abs/1809.01610. arXiv: 1809.01610.

Rimma Pivovarov and Noemie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 22, 04 2015.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. November 2017. URL http://arxiv.org/abs/1711.05225.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. 2018. URL https://arxiv.org/abs/1804.07839.

Thomas Schlegl, Sebastian M Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, January 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-018-0300-7. URL http://www.nature.com/articles/s41591-018-0300-7.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2015. URL http://arxiv.org/abs/1411.4555.

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. Deep learning for identifying metastatic breast cancer. *CVPR*, 2016. URL https://arxiv.org/abs/1606.05718.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. pages 2097–2106, 2017. URL http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html.

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. January 2018. URL http://arxiv.org/abs/1801.04334.

R. J. Williams and D. Zipser. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280, June 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.2.270.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. *CoRR*, 2016. URL http://arxiv.org/abs/1611.01646.