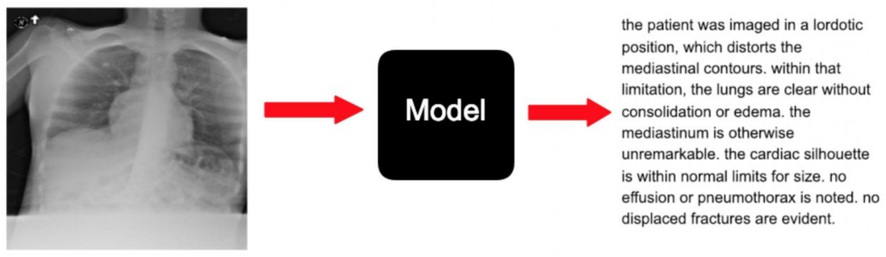
Baselines for Chest X-Ray Report Generation

William Boag[1], Tzu-Ming Harry Hsu[1], Matthew McDermott[1], Gabriela Berner[2], Emily Alsentzer[1], Peter Szolovits[1]

1. MIT CSAIL, 2. Harvard

https://github.com/wboag/cxr-baselines

Generating Radiology Reports



Text Generation could have huge benefits in clinical practice, including:
+ Increasing the speed and consistency of interpretations.
+ Scaling to under-resourced areas (rural America; developing countries).
+ Decrease amount of time documenting patient info in records.
+ Text can be more expressive than structured data (e.g. slight vs mild vs moderate).

Dataset

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

2019 has seen a massive increase in the number of publicly available CXR images.

With all this new data, we need baselines to understand what "good" performance is.

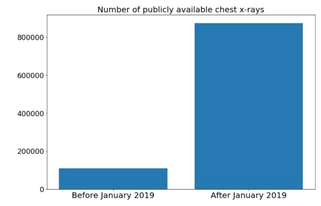


Figure credit: Alistair Johnson

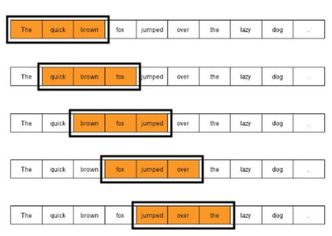
Access the Data: https://mimic-cxr.mit.edu/
Basics from Alistair: https://github.com/mlhc19mit/recitations/blob/master/rec4-slides.pdf

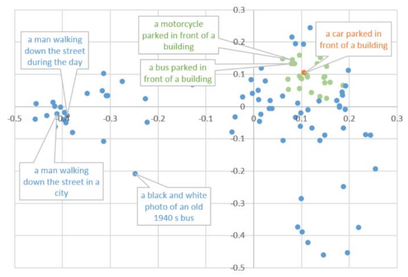
Importance of Baselines

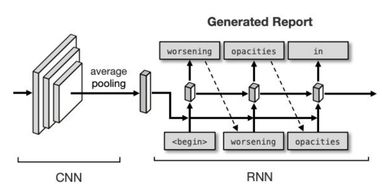
Baselines are essential to progress, and honestly they're just good science. They:
- Diagnose poor performance by isolating different parts of a model.
- Help "debug" the complexity of a dataset.
  - Some authomatically-generated datasets prove to be simpler than expected.
- Ground our understanding in how "hard" a task is.
  - Your model gets 80%. Is that good? How well would a model do by chance?
- Help us test intuition-based hypotheses.
  - We assume the structure of the space looks a certain way.
  - If we probe that, does it behave the way we expect?

Blog Post: tinyurl.com/baseline-blog

Choice of Baselines







Random Train Report
- Irrelevant
+ Grammatical

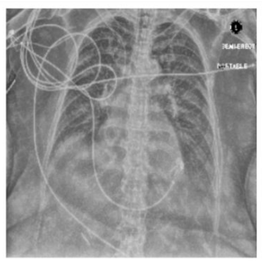
N-grams
+ Simple decoder but clearly relevant
- Ungrammatical

Nearest Neighbor
+ Grammatical
+ Decently relevant
- Might ignore specifics of the particular image.

Figure from https://arxiv.org/pdf/1505.04467.pdf

Show-and-Tell
+ Relevant
+ Most grammatical baseline model

Sample Output

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 .

Evaluation and Results

Natural Language Generation Metrics

Clinical Correctness

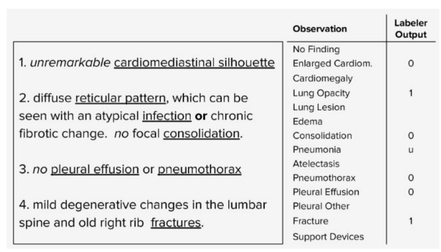


Table 2: Automatic evaluation metrics of baseline methods for image captioning task.

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

Neural Network method (even this simple kind) performed the best.

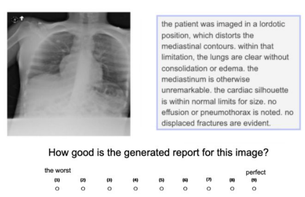
1-NN has decent performance & should be included as a baseline comparison in more generation work.

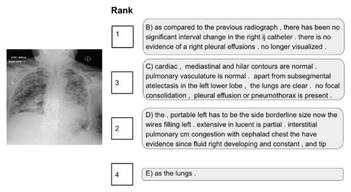
BLEU and CIDEr score Random (irrelevant-but-gramatical) higher than 3-gram (relevant-but-ungrammatical). But 3-gram had higher clinical correctness. That cannot be right. Standard general domain metrics are insufficient.

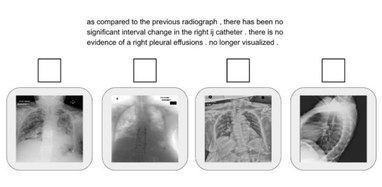
Future Work: Need Better Metrics!

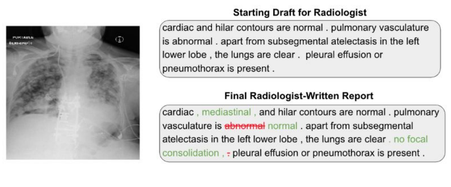
Metrics like CIDEr and BLEU were not validated on clinical data, and they account don't for correctness (just surface-level similarities).

Need to collect clinical judgments from doctors in order to develop new metric which better aligns with "right" thing.









Thanks

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).