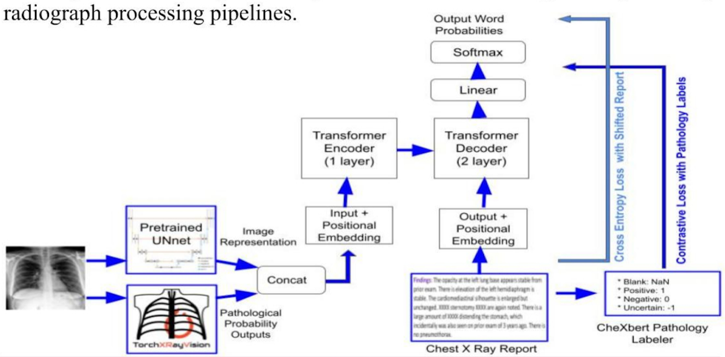
Chest X-Ray Report Generation from Chest-X Ray Images

Esin Darici Haritaoglu | Aleksandr Timashov | Matthew Tan
March 14th, 2022 (Alumni Center in McCaw Hall, CS224n)

Problem

The automatic **generation of** highly clinically accurate **radiology reports** from Chest X-Ray images could **improve clinical outcomes** by reducing radiologist workload, prioritizing severe cases, and augmenting existing radiograph processing pipelines.


Output Word Probabilities
Softmax
Linear
Transformer Encoder (1 layer)
Transformer Decoder (2 layer)
Cross Entropy Loss with Shifted Report
Contrastive Loss with Pathology Labels
Pretrained UNnet
Image Representation
Input + Positional Embedding
Output + Positional Embedding
Concat
TorchXRayVision
Pathological Probability Outputs
Chest X Ray Report
* Blank: NaN
* Positive: 1
* Negative: 0
* Uncertain: -1
CheXbert Pathology Labeler

Findings: The opacity at the left lung base appears stable from prior exam. There is elevation of the left hemidiaphragm is stable. The cardiomediastinal silhouette is enlarged but unchanged. XXXX sternotomy XXXX are again noted. There is a large amount of XXXX distending the stomach, which incidentally was also seen on prior exam of 3 years ago. There is no pneumothorax.

Techniques

- Template matching. It is too restricted method, we did not consider it.
- Retrieval-based. Baseline method. Use K tags from most similar images.
- Encoder-Decoder Generative model;
  There are a lot of things to try. It is our main method.

Takeaways

- Providing **image representation** and **pathological probability** outputs to encoder improves the performance
- **Joint loss** helps significantly

Literature

- **WCL:** Cluster reports with labels for contrastive loss.
- **IFCC:** Combine factual metric loss with a language model loss and an NLG loss.

Metrics

- NLG metric **BLEU** doesn't show Clinical Efficacy (CE).
- Compare pathology labels from original and generated text for **CE metrics**.
- Baseline method uses tag Retrieval from corpus of ground truth clinical tags.

Data & Experiments

- **IU X-Ray** Frontal images, reports and pathology labels (1952 for training, and 488 for testing)
- Experiment with/out pathological probability outputs
- Experiment with/out contrastive loss

Results

Dataset	Best Model from the paper	NLG Metrics		CE Metrics		
		BLEU-1	BLEU-4	Precision	Recall	F1
Mimic-Cxr	IFCC	-	11.1	46.0*	72.9*	56.4*
	WCL	37.3	10.7	38.5**	27.4**	29.4**
	Ours	In progress				
IU X ray	Retrieval	0.78	Percent correct tags generated			
	R2Gen	47.0	16.5			
	CMN	47.5	17.0			
	Ours	27.7	2.0			

*The micro average of accuracy, precision, recall, and F1 scores are calculated over 5 observations for: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion
** It is not explicitly stated but we concluded that WCL results are macro average over all 14 observations - both results use CheXpert (not CheXbert)

Future Work

- Train and validate on a full-size **MIMIC-CXR** dataset. (it is not possible in project time due to limited computational resources)
- Experiment with **model architectures** for pathological probabilities **class predictions**.
- Experiment with different architectures for **generation based approach**.
- Experiment more with **joint loss functions**, including contrastive loss functions;