Simulating Deployment of a Multi-{Modal, Task, Site} Chest X-Ray Model

Cara Van Uden, Anuj Pareek, and Andrew Gaut

CS231n - Deep Learning for Computer Vision

Introduction

- Pneumonia is a major cause of morbidity and mortality worldwide. Patients are typically diagnosed from a combination of chest X-rays (CXRs), clinical symptoms, and blood tests.
- If we can diagnose pneumonia from CXRs using ML, we can help patients receive better and faster treatment by:
  - reducing radiologist workload,
  - reducing diagnosis cost, and
  - accelerating time to diagnosis.
- We collaborate with Intermountain Healthcare to update and improve CheXED, their currently deployed pneumonia detection model.

Problem Statement

- We leverage **real-world multimodal data (CXRs and radiology reports)** from collaborators at Intermountain Healthcare to develop an **image+text-SSL multi-task model** for pneumonia, multilobar pneumonia, and pleural effusion.
- We **simulate deploying** this model in the healthcare system's emergency departments in 2020, and **"retrain"** it with a regular cadence through 2021.
- We perform a detailed **error analysis** and **demographic subgroup performance analysis** to prepare our model for deployment.

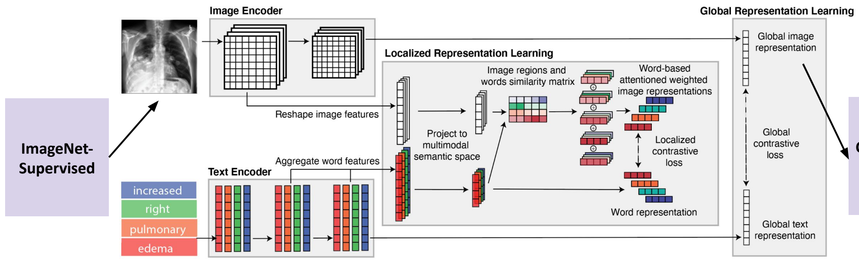
Dataset

- **CheXpert**: 224,316 CXRs and accompanying radiology reports from Stanford Medical Center from 2002-2017.
- **Intermountain**: 16,708 CXRs and accompanying radiology reports from Intermountain Emergency Departments and Urgent Cares from 2009-2021.
  - **Train/val/test**: 15,774/467/467.
- **Preprocessing**: Random image crops and normalization.

Statistics	Train	Validation	Test
CXR Studies, N	15774	467	467
Years, N (%)			
2009-2015	6551 (41.5)	-	-
2020	5490 (34.8)	-	-
2021	3733 (23.7)	467 (100)	467 (100)
Medical Facility			
Emergency	11081 (70.2)	467 (100)	467 (100)
Urgent Care	4693 (29.8)	-	-
Labels, N (%)			
Pneumonia			
Positive	5372 (34.1)	70 (15.0)	71 (15.2)
Uncertain	3872 (24.5)	86 (18.4)	82 (17.6)
Negative	6530 (41.4)	311 (66.6)	314 (67.2)
No. of Lobes			
Single Lobe	5381 (58.2)	42 (26.9)	48 (31.4)
Multilobar	3773 (40.8)	107 (68.6)	98 (64.1)
Unknown	90 (1.0)	7 (4.5)	7 (4.6)
Pleural Effusion			
Positive	1390 (8.8)	24 (5.1)	20 (4.3)
Negative	14384 (91.2)	443 (94.9)	447 (95.7)

Statistics	2021 Data
Patients, N	4651
Age, mean ± SD	50.8 ±22.8
Female, N (%)	2418 (52.0)
Race, N (%)	
White	3835 (82.5)
Hawaiian or Pacific	144 (3.0)
Black or African American	107 (2.3)
Asian	64 (1.4)
Other or Unavailable	501 (10.8)

Methods and Results



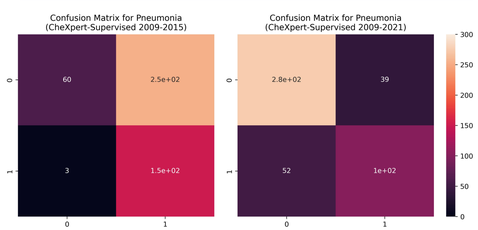
CheXpert-SSL (GLoRIA)

Task	Pretrain Method	Pretrain Dataset	Train Dataset	AUROC	AUPRC	F1 Score
Pneumonia	Supervised	CheXpert	'09-'15	0.801	0.690	0.665
	Supervised	ImageNet	'09-'21	0.860	0.779	0.713
	Supervised	CheXpert	'09-'21	0.861	0.783	0.706
	Supervised	All CXRs	'09-'21	0.852	0.766	0.700
	MoCo	CheXpert	'09-'21	0.863	0.790	0.709
	GLoRIA	CheXpert	'09-'21	0.863	0.795	0.722
	GLoRIA++	CheXpert	'09-'21	**0.880**	**0.810**	**0.738**
Effusion	Supervised	CheXpert	'09-'15	**0.961**	**0.696**	**0.780**
	Supervised	ImageNet	'09-'21	0.954	0.597	0.634
	Supervised	CheXpert	'09-'21	0.953	0.596	0.732
	Supervised	All CXRs	'09-'21	0.926	0.466	0.595
	MoCo	CheXpert	'09-'21	0.947	0.540	0.600
	GLoRIA	CheXpert	'09-'21	0.955	0.589	0.708
	GLoRIA++	CheXpert	'09-'21	0.959	0.647	0.683
NumLobes	Supervised	CheXpert	'09-'15	0.865	0.938	0.851
	Supervised	ImageNet	'09-'21	0.873	0.943	0.869
	Supervised	CheXpert	'09-'21	0.881	0.945	0.869
	Supervised	All CXRs	'09-'21	0.866	0.945	0.867
	MoCo	CheXpert	'09-'21	0.874	0.946	0.866
	GLoRIA	CheXpert	'09-'21	**0.900**	**0.955**	**0.894**
	GLoRIA++	CheXpert	'09-'21	0.883	0.949	0.876

Better performance from GLoRIA++ on Pneumonia and Effusion.

Better performance from GLoRIA on Multilobar Pneumonia (NumLobes).

We notice a data distribution shift between 2015 and 2021. The currently deployed model classifies most examples as positive with its current operating point threshold.



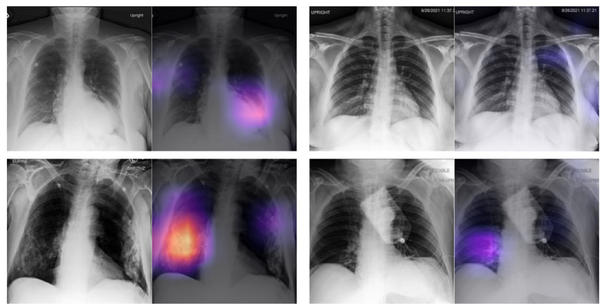
Retraining with a regular cadence leads to improved performance for the pneumonia task.

Task	SRT?	Train Dataset	AUROC	AUPRC	F1 Score
Pn.	No	'09-'15	0.827	0.701	0.678
	No	'09-'20	0.881	0.768	0.726
	Yes	'09-'20	**0.906**	**0.802**	**0.780**
Eff.	No	'09-'15	**0.987**	**0.817**	0.744
	No	'09-'20	**0.987**	0.776	**0.813**
	Yes	'09-'20	0.982	0.712	0.757
M.L.	No	'09-'15	0.864	**0.949**	0.870
	No	'09-'20	**0.886**	0.885	**0.886**
	Yes	'09-'20	0.864	0.945	0.871

Our best model exhibits significant performance discrepancies across several demographic subgroups. This must be corrected before deployment.

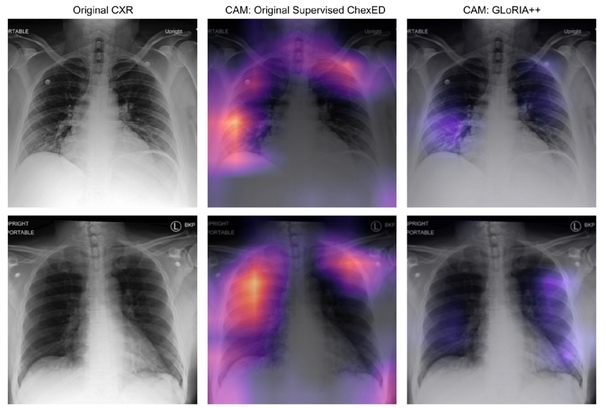
Demographic	Subgroup Name	AUROC	AUPRC	F1-Score
Race	White	0.873	0.801	0.715
	Haw./Pac. Isl.	0.741	0.767	0.800
	Black	0.833	0.900	0.667
	Asian	1.000	1.000	0.889
	Nat. Am.	1.000	1.000	1.000
Latino/Spanish	No	0.868	0.79	0.717
	Yes	0.901	0.88	0.767
Gender	Male	0.887	0.865	0.773
	Female	0.868	0.735	0.680
Age	<19	1.0	1.0	0.5
	19-45	0.924	0.894	0.786
	45-65	0.856	0.855	0.744
	65-80	0.795	0.703	0.632
	80+	0.841	0.801	0.790
Insurance Plan	Medicare	0.795	0.757	0.676
	Medicaid	0.891	0.777	0.681
	Tricare	0.933	0.927	0.750
	Other Ins.	0.944	0.896	0.821
	Self Pay	0.756	0.719	0.750
Hospital Name	IM	0.866	0.799	0.773
	MK	0.898	0.866	0.743
	RV	0.891	0.809	0.681
	AV	0.852	0.694	0.710
	LD	0.910	0.831	0.571

Results



Many false positives and negatives arise from understandable mistakes.

Figure 3. The figure shows selected TP, TN, FP and FN examples of class activation maps for pneumonia using the updated GLoRIA++ model. *Top Left*: True positive. There is a silhouette sign at the left cardiac border as a result of pulmonary consolidation in the left lung lingula segment. The model classifies the example correctly as pneumonia, and the CAMs highlight the abnormality correctly. *Top Right*: True negative. The lung fields are clear bilaterally with no sign of pneumonia. The CAMs do not highlight any focal area consistent with a true negative example. *Bottom Left*: False positive. Extensive pleural calcifications are seen in lung fields bilaterally. The CAMs show that the model incorrectly classifies these calcifications as pneumonia. *Bottom Right*: False negative. There is atelectasis of the left lower lobe with a faint air bronchogram consistent with consolidating pneumonia. The CAMs do not highlight the area with pneumonia, possibly because the model was unable to locate pneumonia-like features within the atelectic lung tissue.



GLoRIA++ is more robust to real-world "hard examples" than supervised CHeXED.

Figure 4. The figure shows examples of hard pneumonia cases which were classified incorrectly by the original CheXED, but classified correctly by GLoRIA++. *Top row*: False positive. A normal chest X-ray acquired suboptimally in the expiratory breathing phase which leads to vessel crowding in basal lung fields. CAMs from original CheXED show that this vessel crowding in the right lung resulted in an incorrect classification of pneumonia, but the GLoRIA++ model correctly classified the chest X-ray as negative for pneumonia and CAMs did not excessively highlight any focal lung areas. *Bottom row*: False positive. A completely normal chest X-ray but with cardiac electrodes on the shoulders. The CAMs from the orignal CheXED highlight the cardiac electrodes which are mistaken for pneumonia. The GLoRIA++ model does not excessively highlight any areas on this normal chest X-ray and classifies it correctly as negative for pneumonia.

Conclusion and Next Steps

- CheXpert-GLoRIA++ CheXED achieves **top or top-2 performance** on all tasks.
- Differences between GLoRIA and GLoRIA++ task performance can be explained by the supervised task set, which indicates **specific use cases for each model**.
- We can **maintain performance in the face of data drift** via regular retraining after deployment. We improve performance for the pneumonia task with this method.
- Qualitative results indicate that GLoRIA++ is more **robust to real-world "hard examples"** than the currently deployed model, and that GLoRIA++'s mistakes are often on CXRs that radiologists would also find more difficult.

- **Next steps**: debias model, retrain with more recently-available data (both labeled and report-only), perform more few-shot and dataset size ablation studies to give annotation recommendations to collaborators.