On the limits of cross-domain generalization in automated X-ray prediction

Joseph Paul Cohen[12], Mohammad Hashir[12], Rupert Brooks[3], and Hadrien Bertrand[1]

[1] Mila, Quebec AI Institute
[2] University of Montreal
[3] Nuance Communications

MIDL Montréal 2020

arxiv.org/abs/2002.02497
github.com/mlmed/torchxrayvision

Initial results when evaluating a model trained on NIH data on an external dataset from Spain.

	Test data (AUC)	
	NIH (Maryland, US)	PadChest (Spain)
Mass	0.88	0.89
Nodule	0.81	0.74
Pneumonia	0.73	0.83
Consolidation	0.82	0.91
Infiltration	0.73	0.60

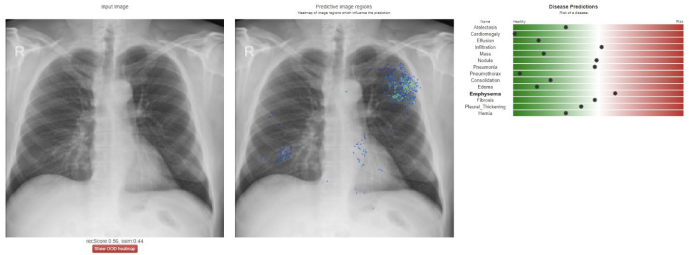
What would lead to such strange results?

An online post about the system indicated some contention about these labels.

Bálint Botz - Evaluating chest x-rays using AI in your browser? — testing Chester, April 2019.

Infiltration, consolidation, pneumonia

Infiltration/consolidation/pneumonia treated as distinct categories feels a bit awkward, as the first two are nonspecific (and largely synonymous) descriptors, while the latter is an actual disease. This categorization has been unfortunately inherited from the NLP-processed training dataset. First I wanted to make this reasonably difficult and selected one of my own cases for this. This time Chester gave an unconvincing result, highlighting an area as suspicious which in my opinion contains no abnormality.



Case courtesy of Dr Balint Botz , Radiopaedia.org, rID: 62068

Many datasets exist with different methods of obtaining labels. Automatic or hand labelled

NIH chest X-ray14
14 labels

Automated rule-based labeler (NegBio)

PADCHEST, ~200 labels

27% hand labelled, others using an RNN.

CheXpert, 13 labels

Custom rule-based labeler.

MIMIC-CXR, 13 labels

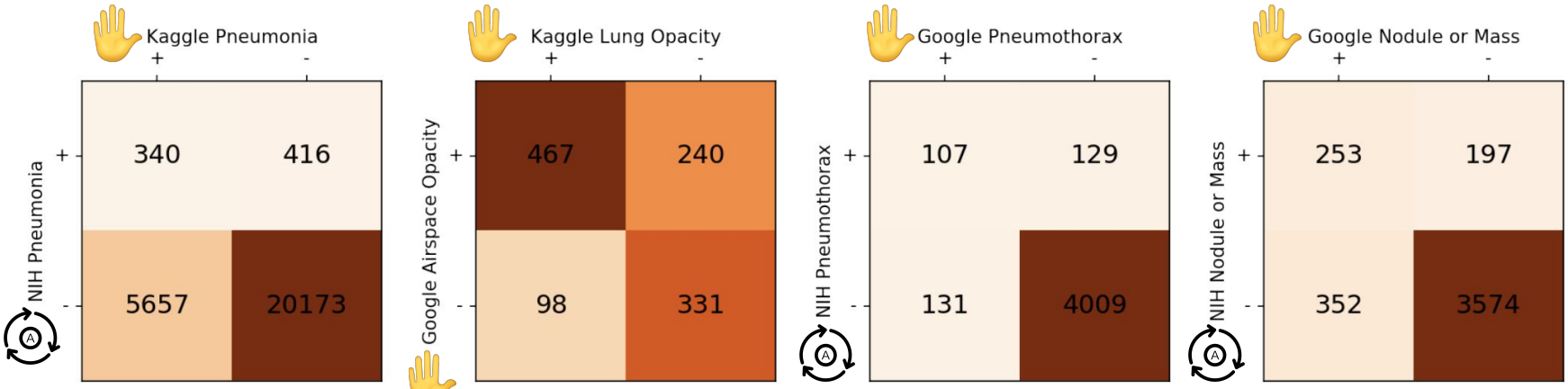
Automated rule-based labeler. NIH (NegBio) and CheX labelers used.

RSNA Pneumonia Kaggle
Relabelled NIH data

A group at Google relabelled a subset of NIH images

MeSH automatic labeller

Label agreement between datasets which relabel NIH images



(a) Pneumonia
F1:10%

(b) Lung Opacity
F1:73%

(c) Pneumothorax
F1:45%

(d) Nodule/Mass
F1:48%

Poor agreement!

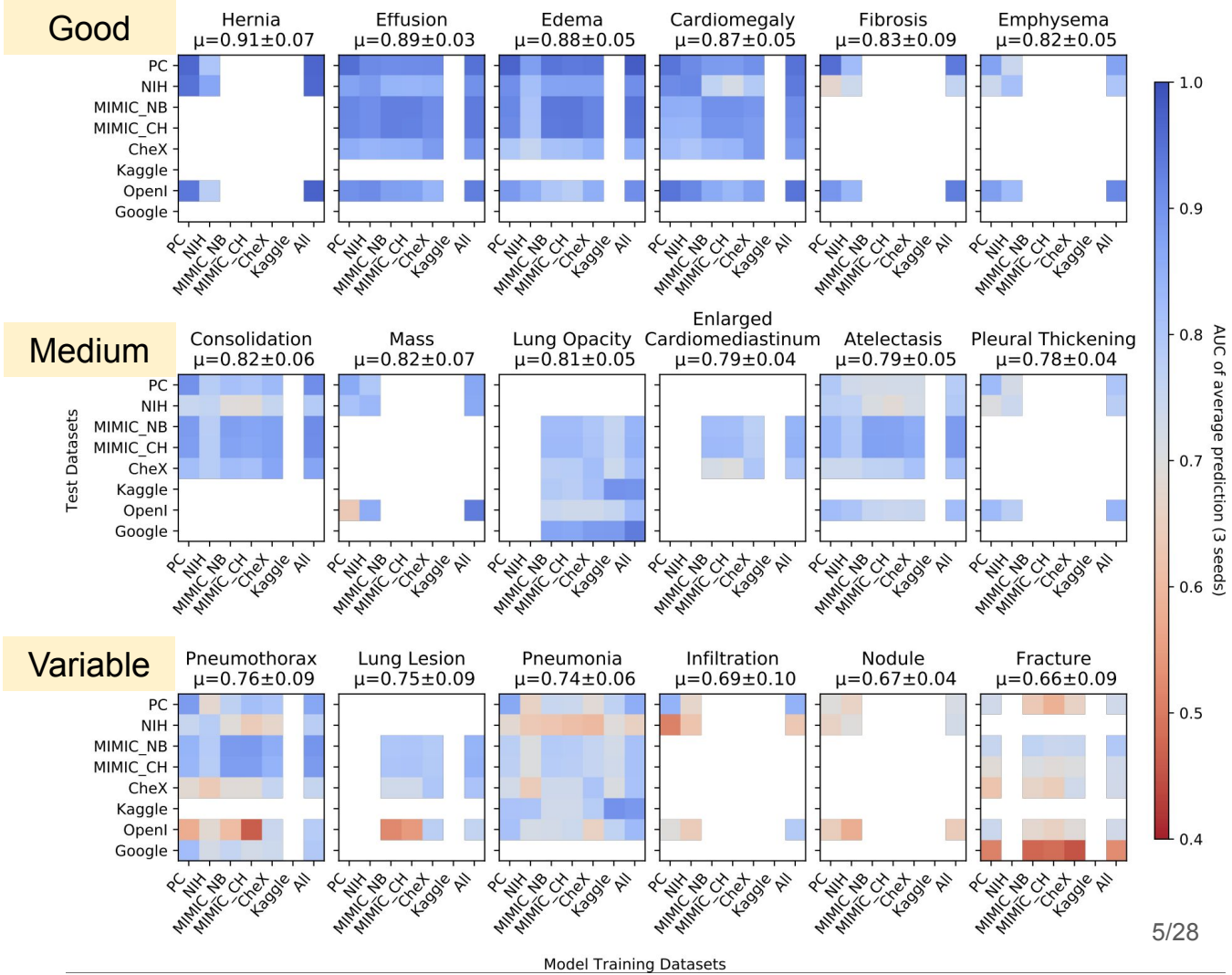
Experiment:

To investigate, a cross domain evaluation is performed. The 5 largest datasets are trained and evaluated on.

Note:

MIMIC_NB and MIMIC_CH only vary based on the automatic labeller.

Task specific agreement!



https://arxiv.org/abs/2002.02497

We model:

$$p(y|x)$$

We may blame poor generalization performance on a shift in x (*covariate shift*) but this would not account why for some y (tasks) it works well.

Possibly reality

$$p(y|x, c)$$

It seems more likely that there is some shift in y (*concept shift*) which would force us to condition the prediction.

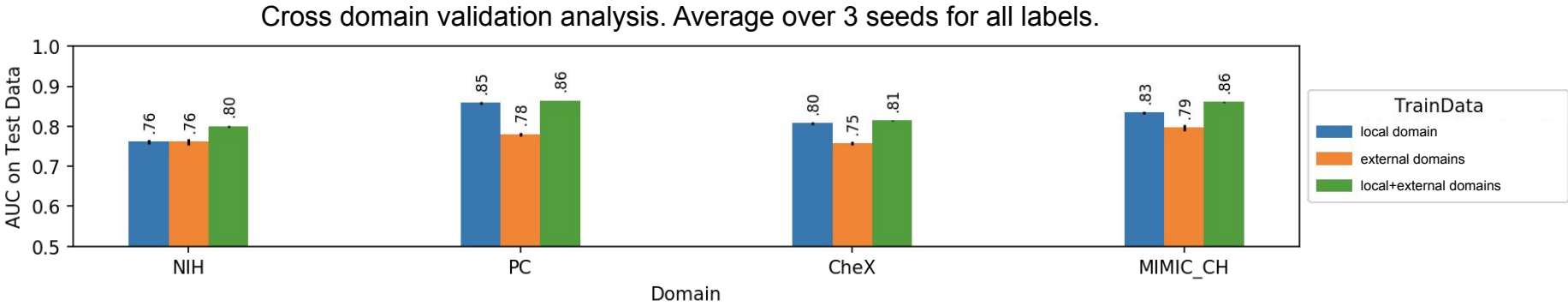
But we want objective predictions!

What is causing this shift?

- Errors in labelling as discussed by Oakden-Rayner (2019) and Majkowska et al. (2019), in part due to automatic labellers.
- Discrepancy between the radiologist's vs clinician's vs automatic labeller's understanding of a radiology report (Brady et al., 2012).
- Bias in clinical practice between doctors and their clinics (Busby et al., 2018) or limitations in objectivity (Cockshott & Park, 1983; Garland, 1949).
- Interobserver variability (Moncada et al., 2011). It can be related to the medicalculture, language, textbooks, or politics. Possibly even conceptually (e.g. footballs between USA and the world). ⚽ 🏈

Are there limits to how well we can generalize for some tasks?

We may think that training on local data is addressing covariate shift

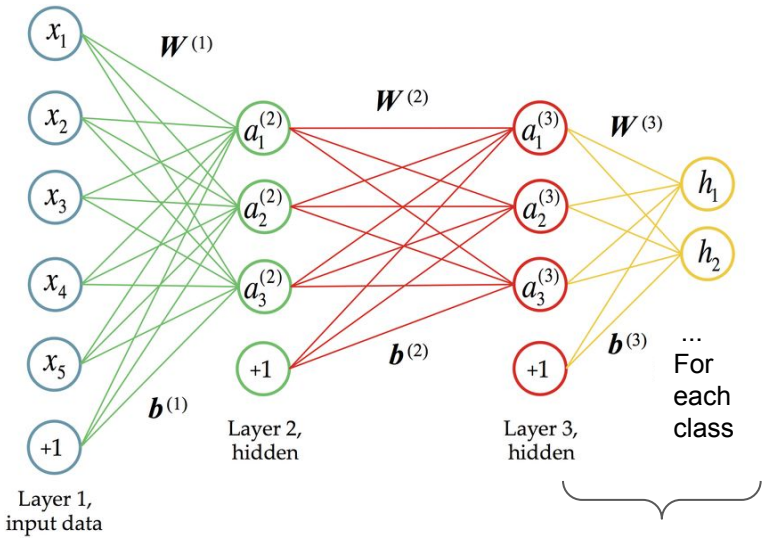


However, training on local data provides better performance than using the larger external datasets.

This may imply the model is only adapting to the local biases in the data which may not match the reality in the images.

How to study concept shift?

We can use the weight vector at the classification layer for a specific task (just a logistic regression)



only this matrix is regularized

$$W \in R^{a \times (t \cdot d)}$$

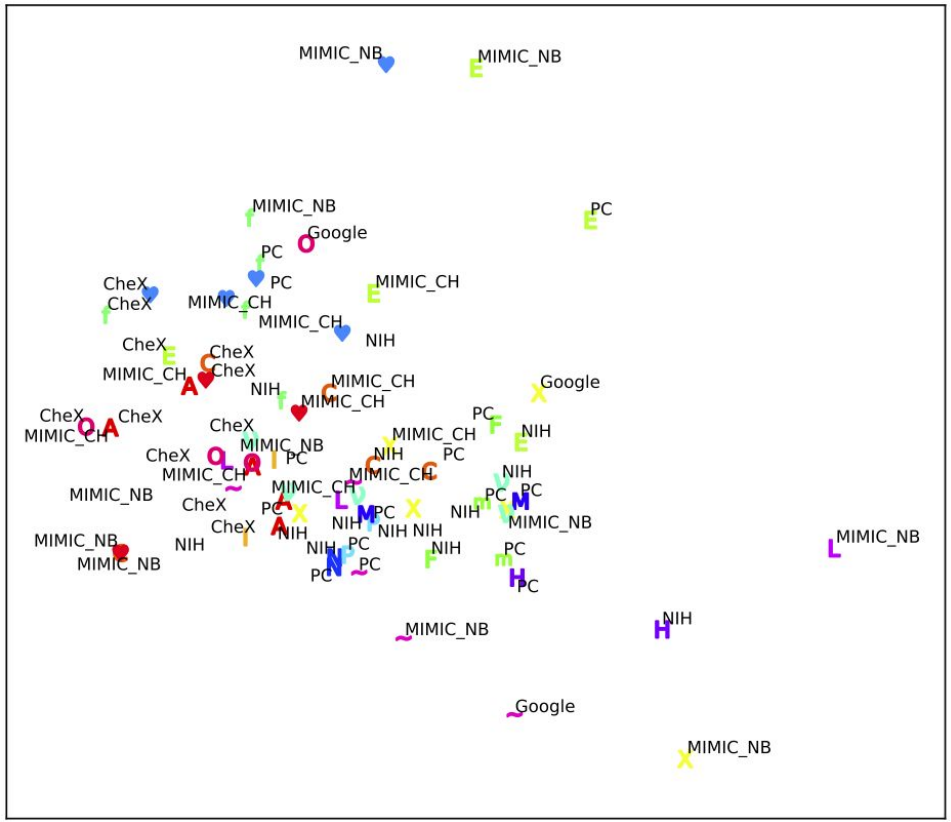
a: feature vector length
t: number of tasks
d: number of domains

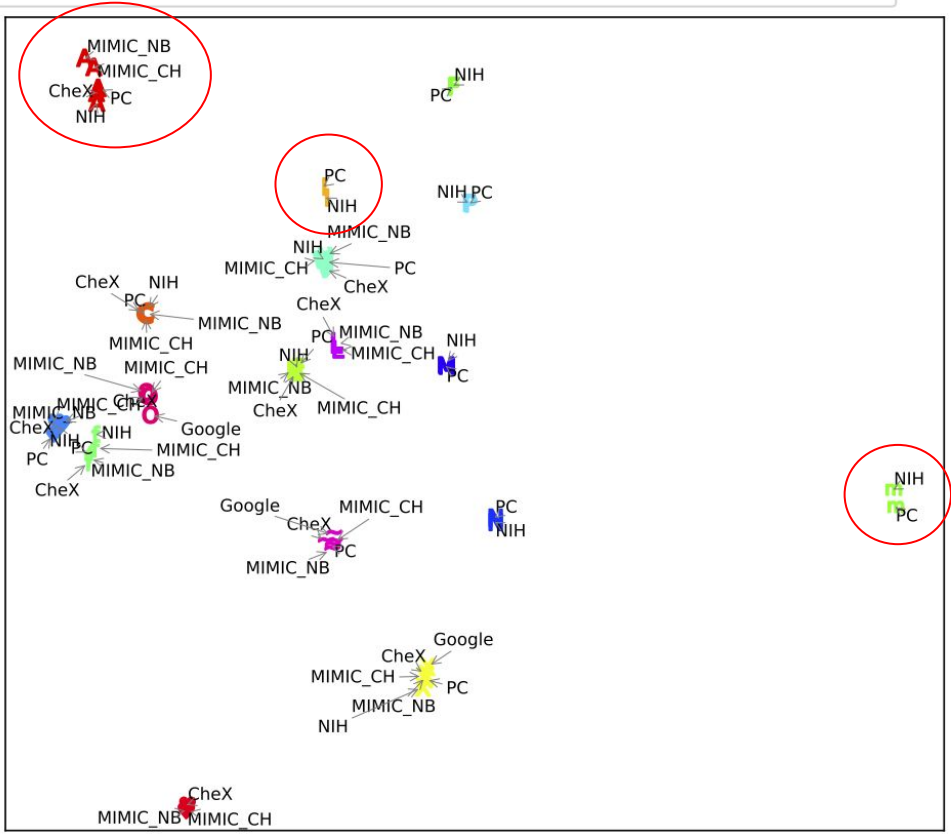
$$||\text{pdist}(W_{t_1}, W_{t_2}, W_{t_3}, \ldots)||_2$$

Minimize pairwise distances between each weight vector of the same task.

If each weight vector doesn't merge together then some concept drift is pulling them apart.

Network figure credit: Sara Sheehan

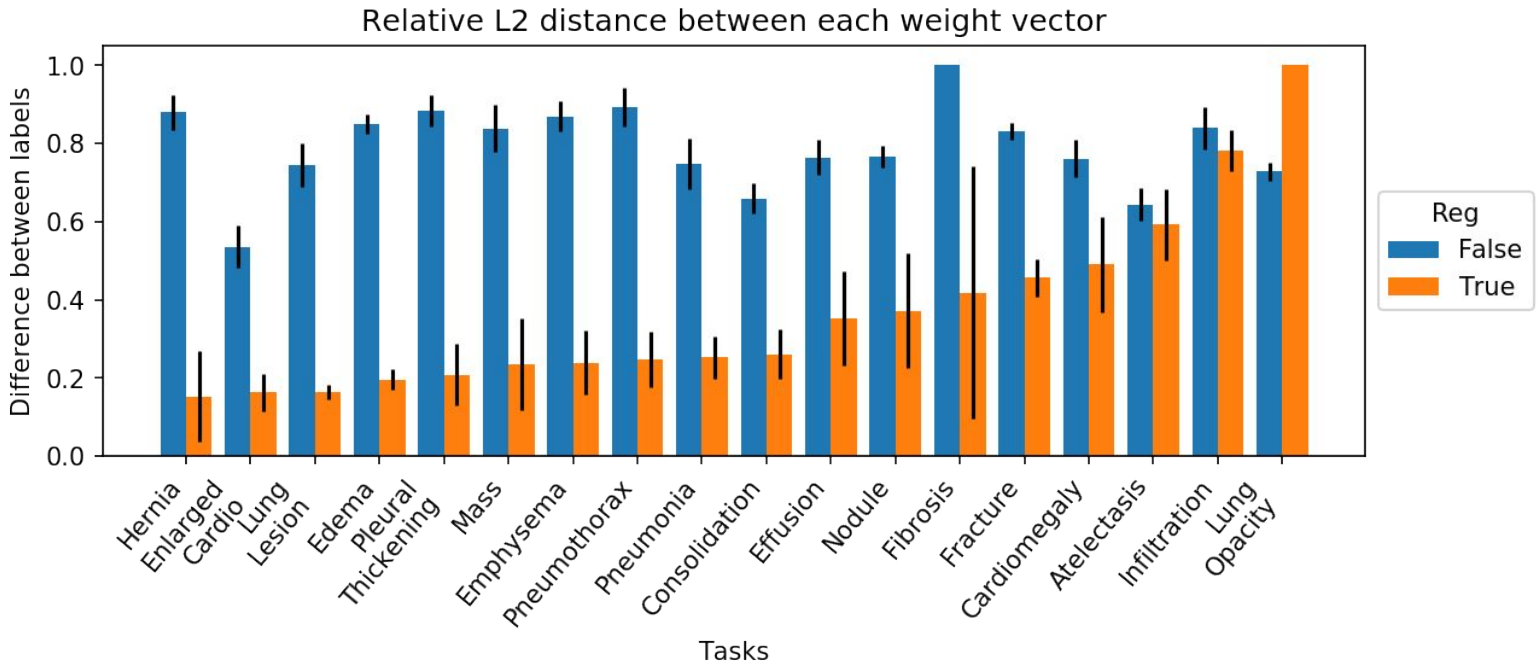




Without regularization

With regularization

Do distances between weight vectors explain anything about generalization?



Sorted based on average distance over 3 seeds some tasks are grouped together easier than others.

Conclusions

- The community may want to focus on concept shift over covariate shift in order to improve generalization.
- Better automatic labeling may not be the answer.
  - General disagreement between radiologists or subjectivity in what is clinically relevant to include in a report.
- We can consider each task prediction as defined by its training data such as "NIH Pneumonia" or "CheXpert Edema" each possibly providing a unique biomarker. The output of multiple models can be presented to a user.
- It does not seem like a solution to train on a local data from a hospital.

Thanks!

arxiv.org/abs/2002.02497
github.com/mlmed/torchxrayvision