Generalization challenges and making models right for the right reasons in medicine
(with a focus on chest X-ray diagnostics)

Joseph Paul Cohen
Postdoctoral Fellow
Mila, Université de Montréal

Mila Medical
Academic Division

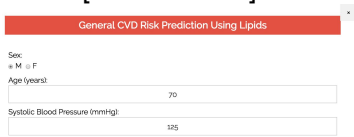
Conflicts of Interest

None

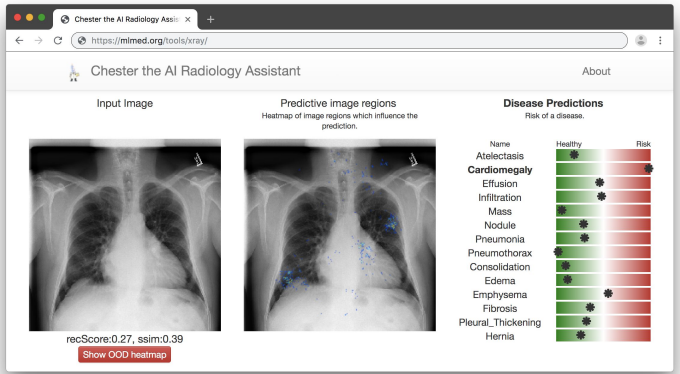
Chester: a free open source tool to try deep learning

Kidney Donor Risk Index (KDRI) [Feng 2006]

The Framingham Heart Study Cardiovascular Disease Risk [Online ~2018]



Chester: AI Radiology Assistant [Cohen 2019]



NOT FOR MEDICAL USE YET

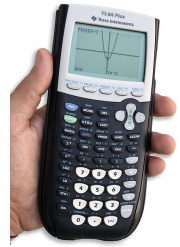
Emergency Room (Time limited human)

Rural Hospital (no radiologist nearby)

Triage of cases by non-expert

Number	Name	Colour	Max time
1	Immediate resuscitation	Red	0 minutes
2	Very urgent	Orange	10 minutes
3	Urgent	Yellow	60 minutes
4	Standard	Green	120 minutes
5	Non-urgent	Blue	240 minutes

As an educational tool in school



[Cohen, Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System, 2019] https://arxiv.org/abs/1901.11210

Chapter 1

Cross-domain generalization

Initial results when evaluating this model 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

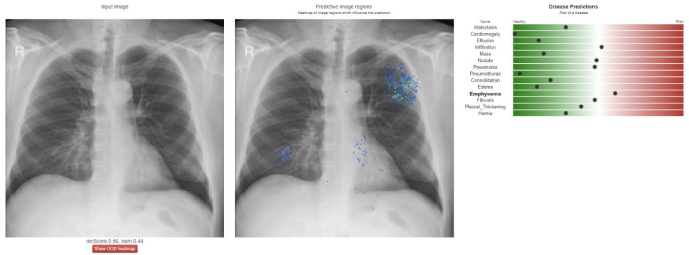
What would lead to such strange results?

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:

Infiltration, consolidation, pneumonia

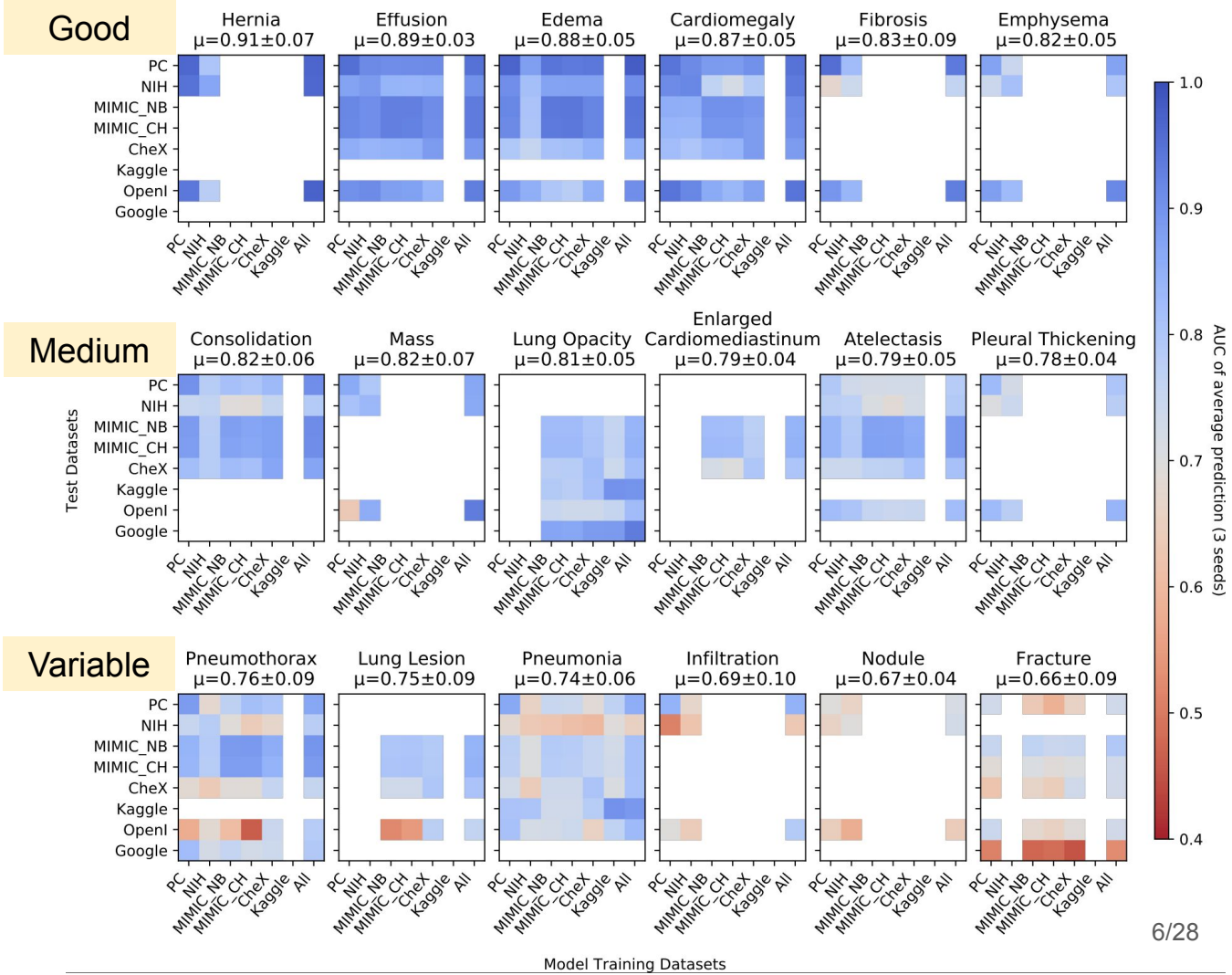
Infiltration/consolidation/pneumonia treated as distinct categories feels a bit awkward, as the first two are nonspecific (and largely synonymous) descriptors, while the latter is an actual disease. This categorization has been unfortunately inherited from the NLP-processed training dataset. First I 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

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

Each dataset's labels are generated using a different method. Some automatic and some manual.



https://arxiv.org/abs/2002.02497

We model:

$$p(y|x)$$

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

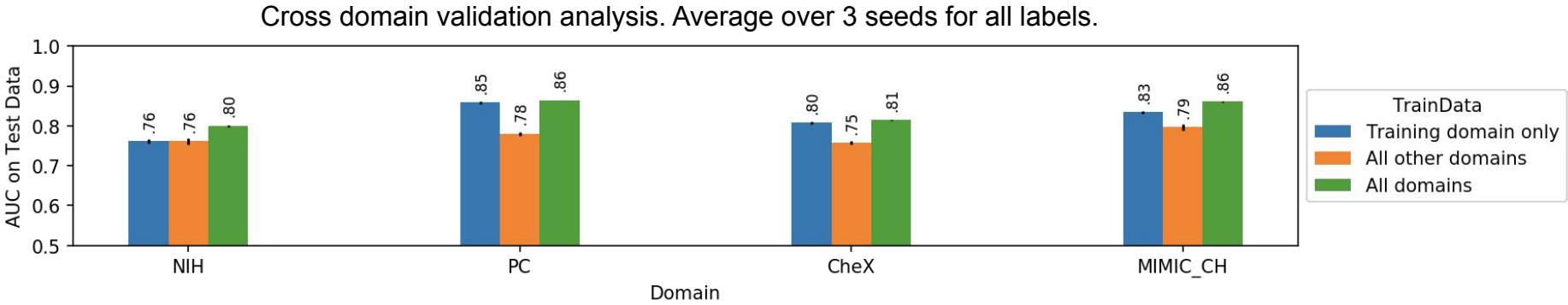
Possibly reality

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

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!

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



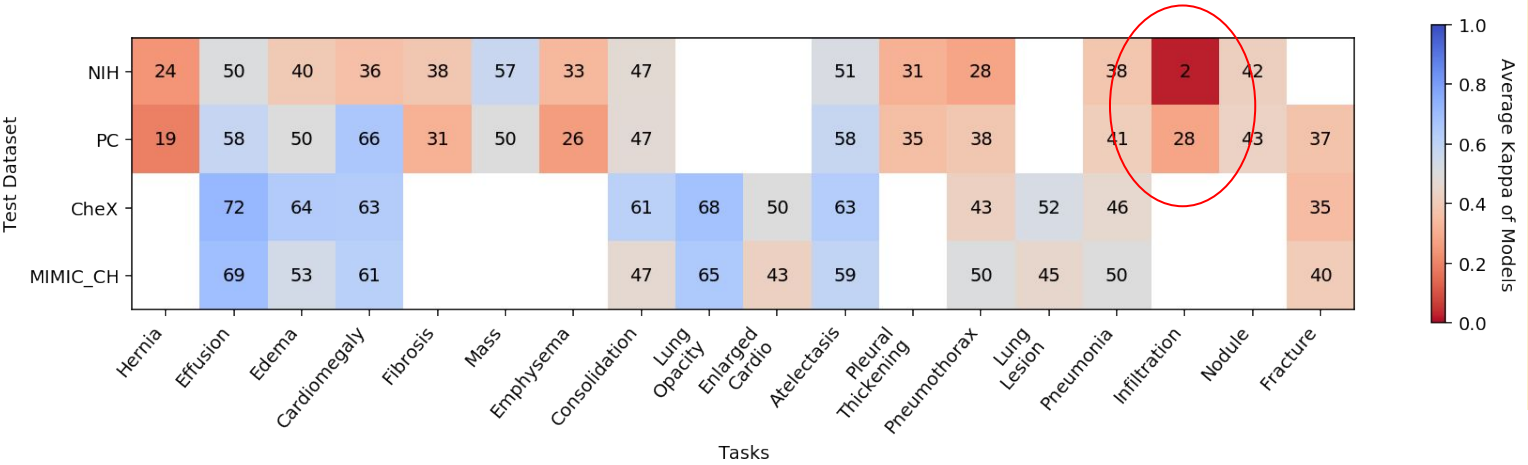
However, training on local data provides better performance than using all other data (>100k examples).

Likely only adapting to the local biases in the data which may not match the reality in the images

What is causing this shift?

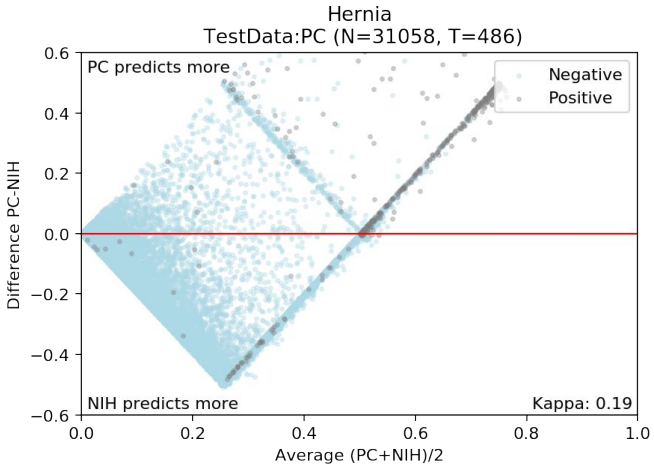
- Errors in labelling as discussed by Oakden-Rayner (2019) and Majkowska et al. (2019), in part due to automatic labellers.
- Discrepancy between the radiologist's vs clinician's vs automatic labeller's understanding of a radiology report (Brady et al., 2012).
- Bias in clinical practice between doctors 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). ⚽ 🏈

Average Kappa between models on a specific dataset. Sorted by generalization accuracy



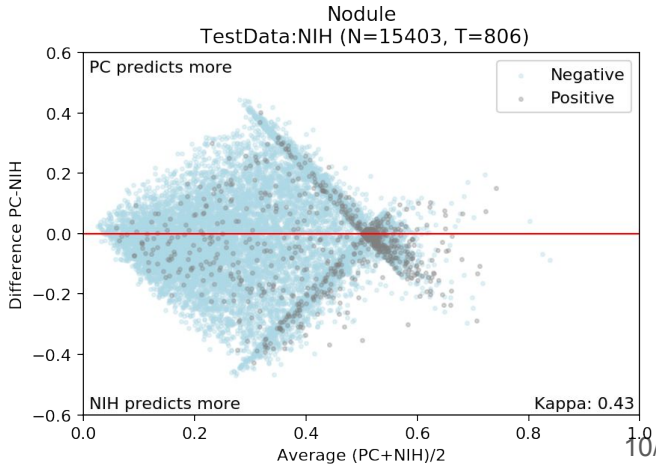
Common labels provide more consistency.

Are labels omitted because they are subject to a lot of interrater variability?



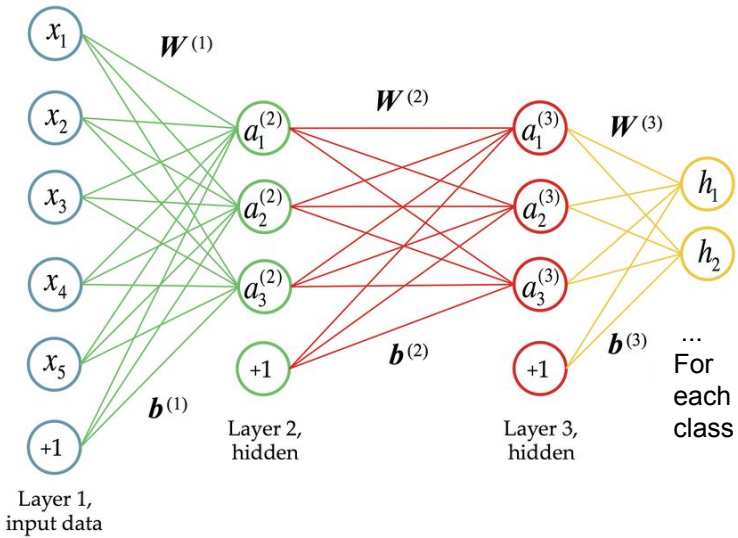
High AUC
Low Kappa

Low AUC
High Kappa



How to study concept drift?

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



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

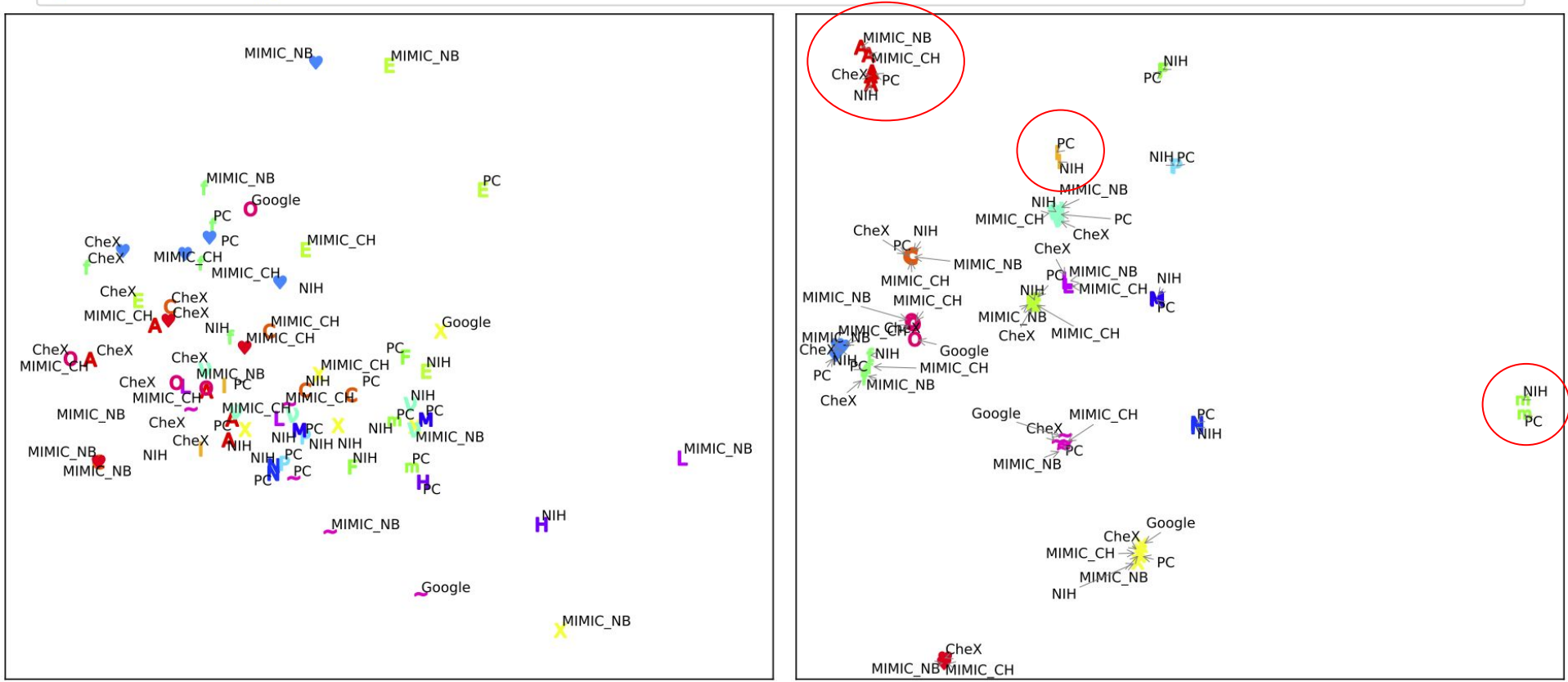
a: feature vector length
t: number of tasks
d: number of domains

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

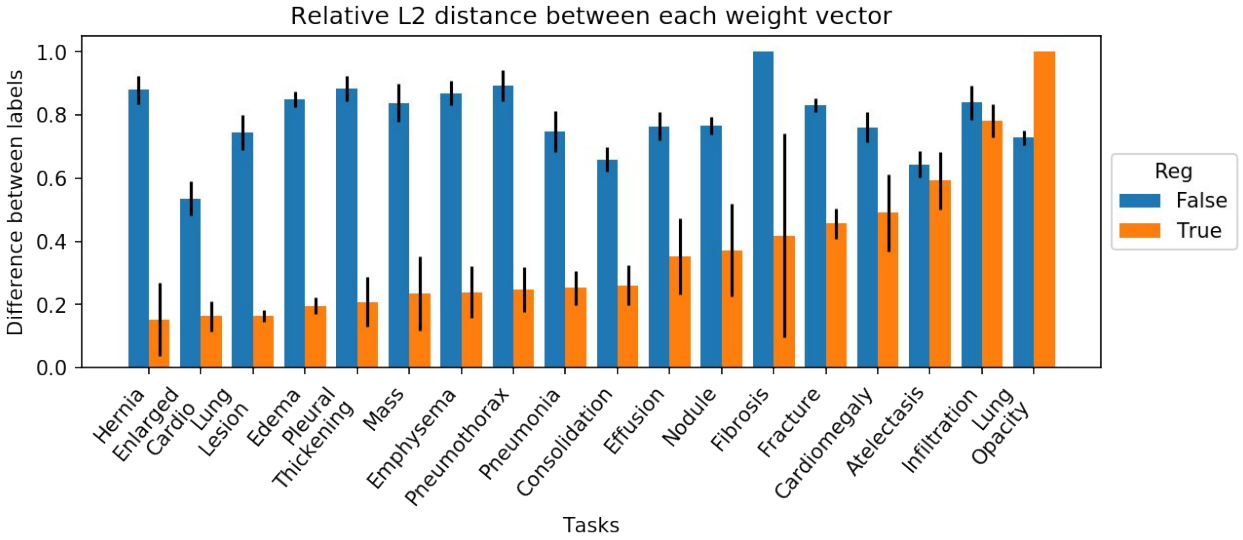
Minimize pairwise distances between each weight vector of the same task.

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

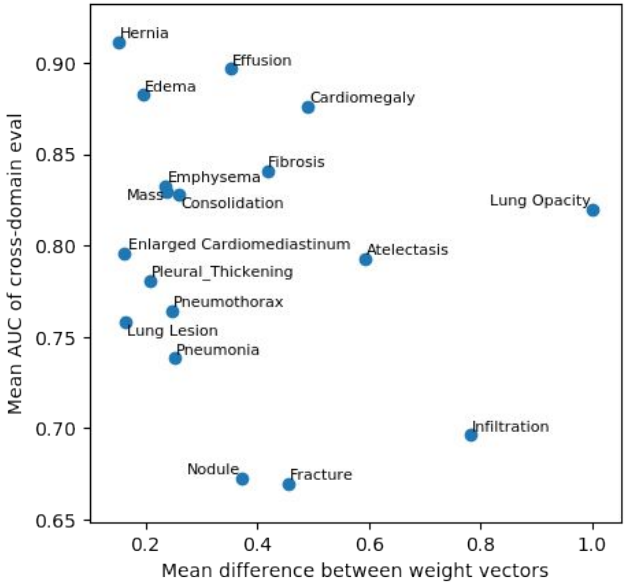
Network figure credit: Sara Sheehan


A Atelectasis
C Consolidation
I Infiltration
X Pneumothorax
E Edema
m Emphysema
F Fibrosis
f Effusion
ν Pneumonia
P Pleural Thickening
♥ Cardiomegaly
N Nodule
M Mass
H Hernia
L Lung Lesion
~ Fracture
O Lung Opacity
♥ Enlarged Cardiomediastinum
MIMIC_NB
MIMIC_CH
CheX
NIH
PC
Google

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.



This distance plotted against average generalization performance shows a slight trend.

Discussion

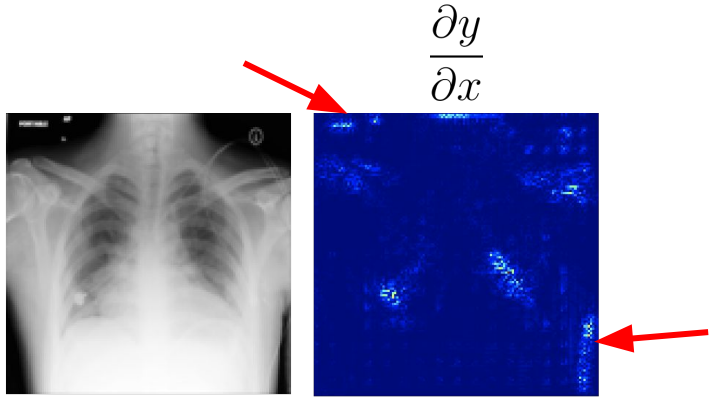
- We believe generalization is not due to a shift in the images but instead a shift in the labels.
- Better automatic labeling may not be the solution.
- General disagreement between radiologists and subjectivity in what is clinically relevant to include in a report.
- We should consider each task prediction as defined by its training data such as "NIH Pneumonia". One can present the output of multiple models to a user.
- We assert that a solution is not to train on a local data from a hospital.

Chapter 2

Incorrect feature attribution

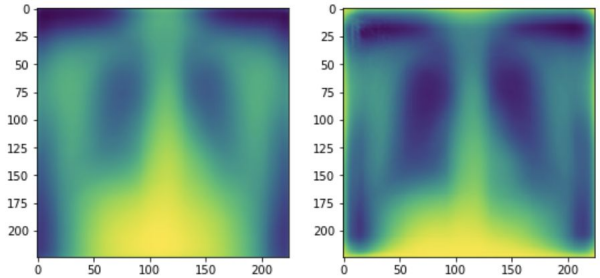
Incorrect feature attribution

Models can overfit to confounding variables in the data.



Overfitting while predicting Emphysema [Vivano 2019]

Example: Systematic discrepancy between
average image in datasets



NIH PADCHEST

- Merging datasets with different class imbalance (confounding artifacts from each hospital)
- Labels confounding with each other
- Demographics confounding with labels

[Zeck, Confounding variables can degrade generalization performance of radiological deep learning models, 2018]
[Viviano, Underwhelming Generalization Improvements From Controlling Feature Attribution, 2019]
[Simpson, GradMask: Reduce Overfitting by Regularizing Saliency, 2019]
[Ross, Right for the Right Reasons, 2017]

Mitigation approaches

Feature engineering

- **Range normalization** (/max)
- **Subspace alignment** (align data using their eigenbasis based on a feature)

During training

- **Reverse gradient** (make intermediate layer invariant to a label) [Ganin & Lempitsky, 2014]
- **Right for the Right Reasons** (regularize saliency map) [Ross, Hughes, & Finale Doshi-Velez, 2017]
- **GradMask** (regularize contrast saliency map between classes) [Simpson, 2019]
- **ActivDiff** (regularize representation to focus on pathology) [Viviano, 2019]

What if feature artifact is correlated with target label?
Is the reason that should be used for prediction known?
What if it is not known?

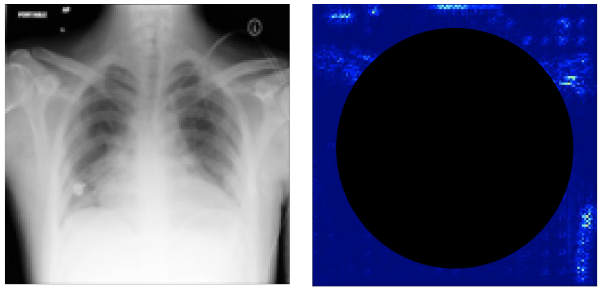
Right for the Right Reasons loss

$$\mathcal{L}_{rrr} = \sum_{\mathbf{x} \in D} \left| \frac{\partial \log \hat{y}}{\partial \mathbf{x}} \cdot (1 - \mathbf{x}_{seg}) \right|_2$$

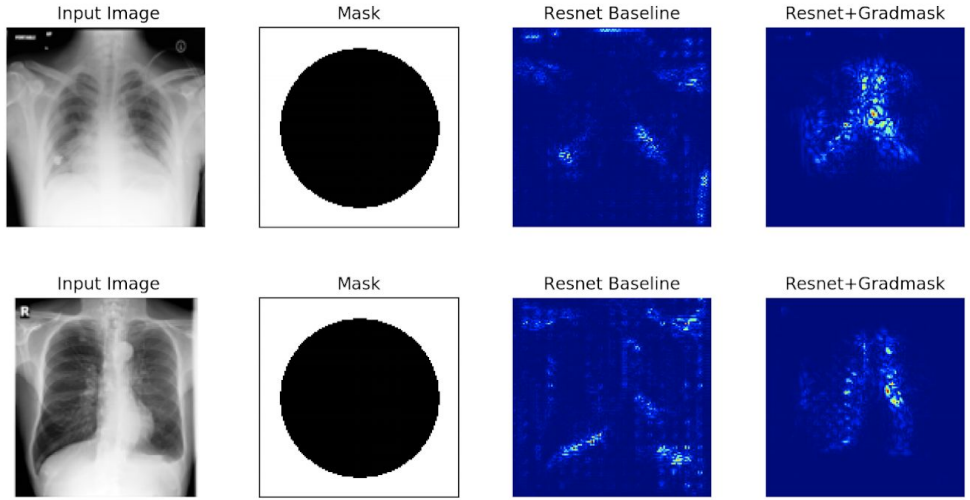
GradMask Contrast loss

$$\mathcal{L}_{gm} = \sum_{\mathbf{x} \in D} \left| \frac{\partial \left| \hat{y}_1 - \hat{y}_0 \right|}{\partial \mathbf{x}} \cdot (1 - \mathbf{x}_{seg}) \right|_2$$

$$\frac{\partial y}{\partial x}$$



Task: emphysema prediction



Experiment Name	AUC
Classify No SPC	$\mathbf{0.70 \pm 0.02}$
Classify w/ SPC	0.44 ± 0.08
Gradmask w/ SPC	0.48 ± 0.03

SPC=site-pathology correlation.

Although the saliancy mask appears more correct the model does not improve.

Joseph Paul Cohen, PhD
Medical Research Lead

MLMED
Machine Learning and Medicine Lab

Mila Medical
Academic Division

Pr. Yoshua Bengio, PhD

Mandana Samiei

Francis Dutil

Martin Weiss

Shawn Tan

Geneviève Boucher

Hashir Khan

Georgy Derevyanko, PhD

Becks Simpson

Hadrien Bertrand, PhD

Karsten Roth

Tristan Sylvain

Margaux Luck, PhD

Sina Honari

Assya Trofimov

Paul Bertin

Tobias Wuerfl

Vincent Frappier, PhD

Joseph Viviano

End