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
Who am I?

Medical Research Lead at Mila
Postdoc at Mila, University of Montreal with Prof. Yoshua Bengio
Director of the Institute for Reproducible Research (U.S. NPO)
Previous: Fellow at Harvard University Herbarium with Prof. Donald H. Pfister
Previous: PhD from University of Massachusetts Boston with Prof. Wei Ding

Satellite imagery analysis

Computer science/ Cyber security education

Neural network training (RandomOut: The Filter Lottery)

Critique of medical image translation

Counting with neural networks

Bluetooth ad-hoc communication

\[ CGN(k) = \sum_i \left| \frac{\partial L}{\partial w_i^k} \right| \]
Who am I?

Medical Research Lead at Mila
Postdoc at Mila, University of Montreal with Prof. Yoshua Bengio
Director of the Institute for Reproducible Research (U.S. NPO)

Previous: Fellow at Harvard University Herbarium with Prof. Donald H. Pfister
Previous: PhD from University of Massachusetts Boston with Prof. Wei Ding

Academic Torrents
(2013)
214k visits/month
53.75TB data total
4TB delivered everyday

ShortScience.org
(2016)
1,437 public summaries
6,400 visits/month

BlindTool
(2015)
27k installs
620 active users (Google stats)
Conflicts of Interest

None
Research program: To create algorithms/tools that are used. To overcome the challenges that prevent mass deployment of AI assistants in healthcare.

- Making medical ML "hip" in the ML community
  - Easing data access and creating public datasets
  - Instead of go-to datasets serving tech giants

- Fundamental challenges in ML that enable medical applications
  - Dealing with generalization issues (out of distribution detection)
  - Making models right for the right reasons (using tumor features to predict a tumor)
  - Self-supervised representation learning for biological modalities (CT, ECG, RNA-Seq)

- Creating/Evaluating ML tools for clinical use
  - Perception augmentation
  - View into historical statistics
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]

Emergency Room (Time limited human)

Rural Hospital (no radiologist nearby)

Chester: AI Radiology Assistant [Cohen 2019]

Triage of cases by non-expert

As an educational tool in school

*NOT FOR MEDICAL USE YET*

Chapter 1
Cross-domain generalization
What would lead to such strange results?

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

<table>
<thead>
<tr>
<th></th>
<th>Test data (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIH (Maryland, US)</td>
</tr>
<tr>
<td>Mass</td>
<td>0.88</td>
</tr>
<tr>
<td>Nodule</td>
<td>0.81</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>0.73</td>
</tr>
<tr>
<td>Consolidation</td>
<td>0.82</td>
</tr>
<tr>
<td>Infiltration</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Initial results when evaluating this model on an external dataset from Spain.

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 Bálint 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.
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 medical culture, 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?
How to study concept drift?

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

\[ W \in \mathbb{R}^{a \times (t \cdot d)} \]

- \( a \): feature vector length
- \( t \): number of tasks
- \( d \): number of domains

\[ \| \text{pdist}(W_{t1}, W_{t2}, W_{t3}, \ldots) \|_2 \]

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.
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.

Example: Systematic discrepancy between average image in datasets

- 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** \((/\text{max})\)
- **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

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

GradMask Contrast loss

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

![Gradient Image](image1.png)  
![Gradient Image](image2.png)
Task: emphysema prediction

Although the saliency mask appears more correct, the model does not improve.

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classify No SPC</td>
<td>0.70 ± 0.02</td>
</tr>
<tr>
<td>Classify w/ SPC</td>
<td>0.44 ± 0.08</td>
</tr>
<tr>
<td>Gradmask w/ SPC</td>
<td>0.48 ± 0.03</td>
</tr>
</tbody>
</table>

SPC = site-pathology correlation.
End