2nd semi-annual report

Oz Sam Kilim ozkilim@hotmail.co.uk
Physics PhD.
Supervisors: Prof. Csabai Istvan, Prof. Pollner Peter.

Domain generalization and representation learning with mammograms

4 projects undertaken during the 2022 spring semester.

1. Physical imaging parameter variation drives domain shift.

Statistical learning algorithms strongly rely on an oversimplified assumption for optimal performance, that is, source (training) and target (testing) data are independent and identically distributed (I.I.D). Variation in human tissue, physician labeling and physical imaging parameters (PIPs) in the generative process, yield medical image datasets with statistics that render this central assumption false. When deploying models, new examples are often out of distribution (O.O.D) with respect to training data, thus, training robust dependable and predictive models is still a challenge in medical imaging with significant accuracy drops common for deployed models. This statistical variation between training and testing data is referred to as domain shift (DS). This work provides evidence that variation in PIPs between test and train medical image datasets is a significant driver of DS and model generalization error is correlated with this variance. We controlled for population shift, prevalence shift, data selection biases and annotation biases to investigate the sole effect of the physical generation process on model generalization for a proxy task of age group estimation on a combined 44k image mammogram dataset collected from five hospitals; Janos J, Nyiregyhaza N, Kecskemét K, CBIS-DDSM and CMMD. The figure below illustrates our findings.

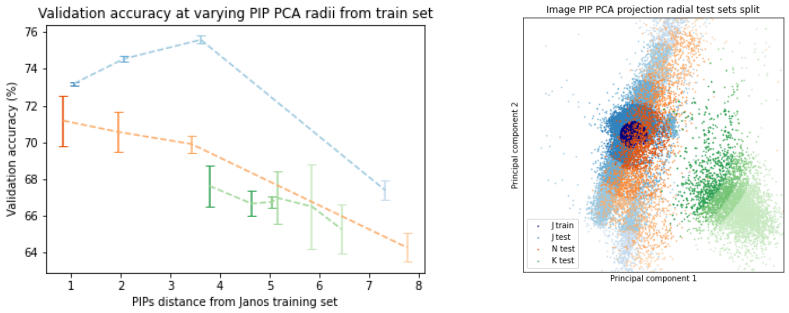

Figure 1. A.Radial experiments where training and testing groups of images are specifically sampled to understand the effect of distance in PIP space with resnet50 classification accuracy for the task of age group prediction. Taking images exclusively from the center of the Janos distribution for a training set and testing on test sets at different radii allows for the study of domain shift as a function of image distance from the training cluster PIP space. B."X-Ray PIP-space" : PCA space for two principal components of 31,000 images PIPs for J, N and K Hungarian datasets.

Domain in medical imaging is treated as a set of non-ordinal discrete labels usually referring to hospitals or vendors but we provided evidence that domain has some continuity due to its dependance on PIPs which are themselves continuous values. This bridging space is common for all existing mammography imaging devices worldwide and is more granular than a simple domain label. Link to full manuscript: https://www.overleaf.com/5281343795msnzptdzmfmc

2. Logistic regression beats classical regression for noisy data tasks.

The initial experiments done to inspect the domain shifting effects between hospitals in **1.** were regressions to predict patient age with a Resnet50 CNN and a MSE loss function. Successful prediction was not possible. The network could not even converge to the mean age of the distribution (58) while a random forest classifier could do this without access to the images themselves. In an attempt to make the task more simple, a classification task between two groups (over and under 58) with a cross entropy loss function was conceived. In this regime the same model could quickly overfit the data. It seemed interesting that "binning" continuous data made the problem learnable when for plain continuous data it was not. Prof. Csabai had seen this same effect in an astronomy project for predicting redshift from spectra where binning made the problem learnable and we wondered if we had stumbled across a more general effect. Below illustrates experiments to investigate this seemingly general but undocumented effect.

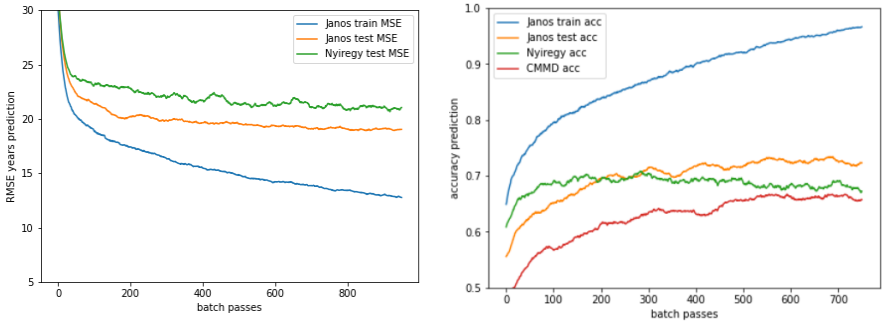

Figure 2. Initial experiments for age prediction from mammogram images with a Resnet50 CNN. A. The failure case where MSE loss decreased over training but tapered at very poor RMSE year predictions for each patient. B. The same task but re-implemented as a classification problem (over or under 58 years of age). The model can quickly overfit and reach a good accuracy on test sets.

To test this in a more controlled environment I used MNIST "4" digits to create a dataset. Images were rotated (uniformly from 0-360 degrees), the new label for each image was the rotation angle applied. Again when binning was implemented (36 degrees for each bin) the same model Resnet50 could achieve better results than the continuous MSE ("Huber regression" in figure 3) loss function trained network when the binning model accuracy was re-cast as a MSE for methodological comparison.

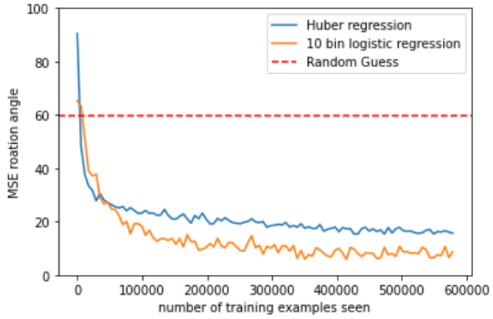

Figure 3. Training of Resnet50 on the task of MNIST 4 images angle rotation prediction. Both continuous methods learn well but the 10 bin logistic regression can achieve better results. This dataset has less noise than the mammography age prediction task.

To inspect this consistently observed effect in the most simple environment possible I implemented a 1 hidden layer dense neural network (no convolutions). The task here was simply to predict a float Y (uniformly sampled from 0-1) given X where X was a 1000 dim long vector of Y's. This task was completely trivial and is quickly learnt to MSE=0 in the continuous regime (seen in blue with Noise "std"=0). With controlled noise added to the X vector we wanted to see how the model performed. As noise increased, the prediction quality decreased. However, for the binned version of the experiment the opposite trend is seen, finally, when the bin number is very large (200) we see a shift back to the continuous learning behavior. This may be some kind of phase transition. Further investigation is needed in the validation metric (potentially to use accuracy) to make sure we are fairly comparing the techniques and not introducing some bias.

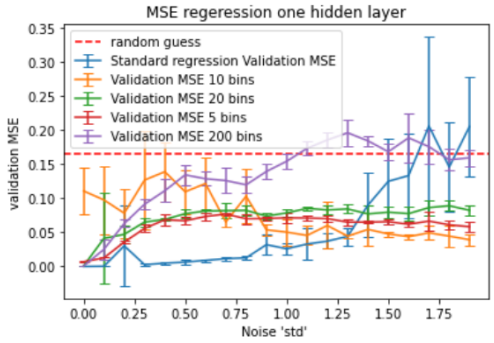

Figure 4. The most simple version of the experiment to compare logistic regression on binned version of ordinal data vs classical regression with a simple 1 hidden layer network as a function of injected noise. This is concordant with the idea that binning seems to increase predictive power of models when the dataset signal to noise ratio is more poor. We theorize that in the limit of many bins, logistic regression reverts back to normal regression in its learning characteristics but this needs more exploration to be confirmed.

3. Mining for long distance tumor biomarkers in mammograms.

This presents the continuation of semester 1 project 1. The experiments were refined further and analyzed. We wanted to see if there were some tumor features in mammograms that are learnable but not visible to humans. To do this we composed a task of distinguishing between a random crop of healthy breast image and a non-tumor containing crop from a tumor containing breast image. Hypothesis: If the network can learn to distinguish them, there is some distant visible signal at a distance from a tumor. We tested this for datasets constructed for variable distances.

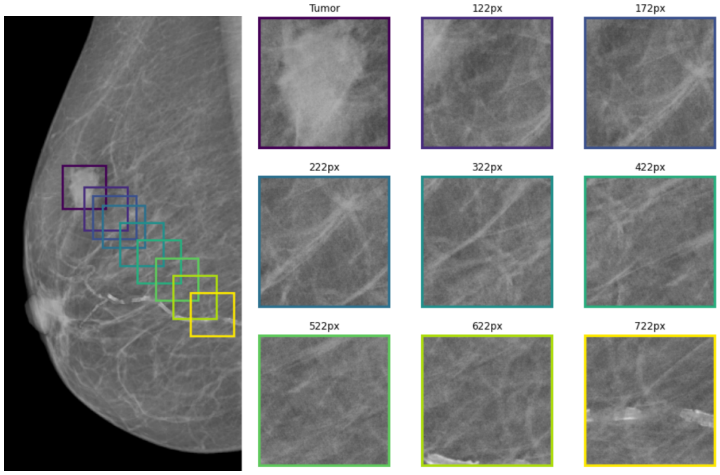

Figure 5. Example of cropping process to create different tissue samples at various distances from the given tumor annotation. We define the distance of a crop D as the distance in pixels between the center of the attomation box and the center of the new crop box.

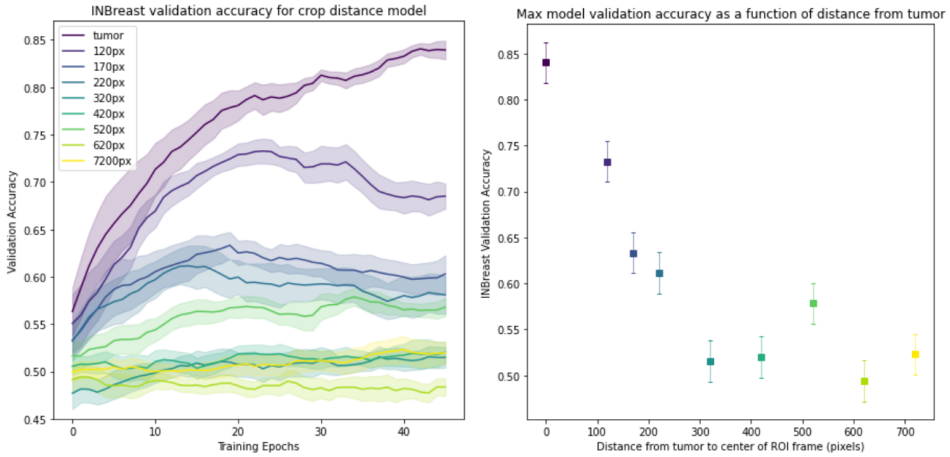

Figure 6. Different non-tumor datasets result. Each curve represents the average of 5 runs with a test set when tested on a Resnet18 network trained to distinguish between the ROI crops and the background tissue crops at a distance from the annotation for the INBreast dataset. There is a small noteworthy bump in model performance for 500px.

https://docs.google.com/document/d/1Mdq07rwCumavdgVZ12Daqg3a5ZHdOHkpIY2jyZK_cDs/edit is a link to the write up of the work including dataset creation, pre-processing and specific experimental details.

4. Covid19 RBD mutations effect on binding affinity.

We have joined a collaborative project with Ákos Gellért and Anikó Mentes to predict binding affinity of various Covid mutants from structures predicted by Alphafold2. We are exploring the use of geometrical deep learning as a potential tool to model surface interactions. We have so far seen a very small correlation between mutant structure variation and experimental binding affinity but we expect chemical and geometrical features will encode more signal with reference to protein-protein binding.

Study activity

1 Module: Deep learning and machine learning in natural sciences FIZ/3/089

Publications

Submitting project 1. To Nature Scientific Reports in the coming weeks.
https://www.nature.com/srep/

Teaching activity

Preparation of kaggle challenge for Deep learning course. Traffic lights sign localization and classification with deep learning.
https://www.kaggle.com/competitions/elte-traffic-signs/overview

Source of the data:https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM
These images are free for inclusion in public reports.