Evaluation and Finetuning of Lung Chest X-ray Models

Rizvan Iskaliev
r.iskaliev@innopolis.university
Innopolis University, BS19-04

5 July 2020

1 Introduction

The goal of this internship is to evaluate the generalization performance of a specific model in Pneumonia detecting when trained and tested on datasets from various institutions that were annotated by different clinicians or labeling tools.

2 Data

In this work the following datasets are used: **Chest X-ray Pneumonia (5k images)**[1]**, RSNA Pneumonia Detection Challenge**[2]**, tb (local dataset from AI Lab), GB7 (local dataset from AI Lab), NIH (Chest 14)**[3]**, all (union of the previous datasets)**
Since all datasets except **NIH (Chest 14)** contain only *Pneumonia* labels, other labels are not considered.

Due to the significant class imbalance and sufficiency of training data, undersampling is employed to the following datasets: **Chest X-ray Pneumonia (5k images), RSNA Pneumonia Detection Challenge, NIH (Chest 14)** datasets. **GB7** dataset has an insufficient quantity of training samples, so loss function with class weights vector initialization is used instead of undersampling.

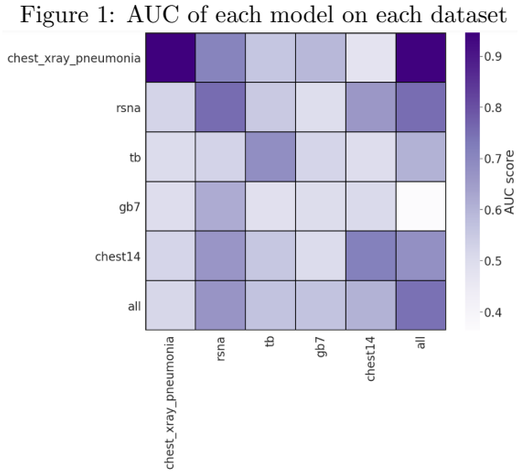
Data augmentation is used to improve generalization: images of the training set are rotated up to 45 degrees, translated up to 10%, and scaled larger or smaller up to 10%. Images of all sets are resized to 224 x 224 pixels, scaled to $[0, 1]$ pixel values and normalized according to *ImageNet's* means and standard deviations.

3 Model and training process

In this work pre-trained on ImageNet dataset *ResNet-50* is evaluated. Data is split into train, validation, and test sets in the ratio $80 : 10 : 10$. The model's layers were frozen except for the last fully-connected layer. *ResNet-50* has been trained for 50 epochs. Adam optimizer with learning rate = 0.001, *CrossEntropyLoss* are chosen as optimizer and loss function respectively. Source code is available on the GitHub repository[4].

4 Evaluation

Figure 1: AUC of each model on each dataset



In Figure 1 a model is trained on each dataset's training subset and then evaluated on the other datasets. If the source dataset and test dataset are the same then the model is evaluated only on the test set, otherwise, the whole dataset is taken for evaluating. AUC is used to determine the performance as datasets have an imbalance in labels.

[1] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[2] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
[3] https://www.kaggle.com/nih-chest-xrays/data
[4] https://github.com/rizvansky/Internship-AI-Lab-IU-2020

5 Discussion

In this work, I verified and found arguments in favor of the results achieved by Cohen et al. (2020): the issue of generalization is because of the shift in the labels in data. These findings explain why the model trained on one dataset does not generalize well on the same data (for example, **GB7**). And even when neural net is trained on all datasets, it does not show decent generalization performance in comparison with other models. In this way, when presenting the prediction of the specific network to a user, additional information about the context and origin of this model should be provided [**1**].

6 Future work

As the main reason for poor generalization performance is the shift in labeling, the achieved results can be improved by enhancement of the labeling process. An important factor is the quality of the training data: if training data consists of many controversial cases then it could negatively affect generalization performance. So, I propose that bounding boxes in X-Ray images for areas that indicate a particular disease should be added during labeling. People with pathologies like Pneumonia tend to have certain symptoms, hence, including some background information about patients could be salutary for making the right prediction.

References

[**1**] J. Cohen, M. Hashir, R. Brooks, H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction", 24-May-2020. [Online]. Available: https://arxiv.org/abs/2002.02497. [Accessed: 06-Jun-2020].