Data Analysis and Presentation Lab (094295)

Final Project: Enhancing X-Ray Imaging Analysis with Active Learning

Nimrod Solomon Matan Shiloni Roni Fridman

Abstract

X-Rays, one of the most used imaging modalities, play a crucial role in diagnosing various medical conditions. However, the effectiveness of machine learning models in this area is often limited by the availability of labeled datasets, which are expensive and time-consuming to produce, as they require expert annotation. Our project focuses on leveraging Active Learning (AL) for the sake of developing a system that prioritizes the annotation of the most informative and challenging X-Ray images, thereby reducing the overall labeling effort required while improving the model's performance. Through experiments utilizing various sampling methods, including a novel "PCA then K-Means" approach, we found that this method consistently equalized traditional techniques in accuracy, F1-score, and precision. Our work demonstrates that Active Learning can achieve comparable accuracy to models trained on entire datasets, even with significantly fewer labeled samples. This project showcases the feasibility of implementing Active Learning in pediatric radiology and highlights opportunities for future research.

Alongside this report, we provide access to the project's GitHub repository for examination, which contains the code developed for the project as well as additional reference materials (refer to the appendices for further details): https://github.com/RoniFridman1/lab_094295

Introduction

Pneumonia is an infection of the respiratory system that primarily impacts the lungs. It arises when harmful microorganisms infiltrate the lung tissue, leading to the destruction of the pulmonary alveoli, which are responsible for oxygen absorption. As a result, these areas become filled with inflammatory fluid, impairing their functionality. Pneumonia can be triggered by a variety of pathogens, including bacteria, viruses, and parasites. While the common symptoms are concentrated around coughing, fever, chills, difficulty breathing, and chest discomfort, in more severe instances, pneumonia may result in complications such as respiratory failure, sepsis, and potentially death. Annually, pneumonia impacts approximately 450 million individuals worldwide and leads to around 4 million fatalities. Although survival rates have been significantly enhanced during the 20th century, pneumonia continues to be a primary cause of mortality in developing nations, as well as among the elderly, infants, and those with chronic health conditions.

One of the most used imaging techniques in radiology, especially for diagnosing conditions affecting the lungs and thoracic region is **chest X-rays**. Interpreting chest X-rays can be challenging due to the complex anatomy and various structures that must be evaluated systematically. In the case of Pneumonia, chest X-rays often reveal opacities in the lungs, which may present as either patchy or confluent areas depending on the extent of infection. Pneumonia typically appears as a consolidation of lung tissue, with visible air bronchograms indicating infection within the airspaces. In radiological practice, detecting pneumonia is crucial as it can lead to complications if untreated. Radiologists evaluate signs such as the silhouette sign, which helps localize lesions, and air bronchograms, which suggest alveolar disease.

Although machine learning may be valuable for radiology-related tasks, such as disease classification from X-ray images, it is often costly to build a dataset for model training. A promising approach for addressing the challenges of data scarcity and annotation costs, including but not limited to the medical domain, is **active learning**. Active learning is a subfield of machine learning and artificial intelligence where the learning algorithm selects the data from which it learns, effectively allowing it to be "curious." By choosing its training data, the algorithm can perform better with less labeled data. Active learning addresses this issue by asking an oracle, often a human annotator, to label only the most informative unlabeled instances. Particularly in the domain of X-Ray imaging, by strategically selecting the most informative images for labeling, active learning algorithms can significantly improve the performance of machine learning models while minimizing the annotation effort required.

In recent years, several active learning techniques have been applied to X-ray image classification tasks with promising results. Those include the "GOAL" strategy, proposed by Nguyen et al. (2021), focusing on selecting the samples for tagging not only based on uncertainty but also on informativeness, derived by clustering. Hao et al. (2021) proposed another data-efficient methodology that can achieve the same level of accuracy as a non-active-learning model using significantly fewer images and labels, based on a Convolutional Neural Network (CNN) unsupervised representation learning, followed by a Gaussian Process (GP) classifier that conducts the active learning pipeline over the learned representations.

While these methods have demonstrated effectiveness in various X-ray image classification tasks, there is still ongoing research to develop even more efficient and robust active learning strategies. Recent advancements in deep learning and generative models have opened new possibilities for active learning, such as using generative models to synthesize new, informative images for labeling. In this project, we explore the application of active learning to the domain of pediatric radiology, specifically focusing on the classification of chest X-rays into "sick" (indicating pneumonia) or "not sick." Our focus is on developing an active learning pipeline that can effectively identify the most informative images for annotation, thereby improving the performance of a pneumonia classification model while minimizing the annotation effort required.

Methodology

Dataset: The dataset we used in this project is Kaggle's "Chest X-Ray Images (Pneumonia)". The dataset consists of 5,863 chest X-ray images of children with and without Pneumonia, performed as part of patients' routine clinical care in Guangzhou, China in 2018. The images are organized into three sets: train, validation, and test, and are categorized into two labels: "Normal" and "Pneumonia". The training set is imbalanced, with 1,341 normal and 3,875 pneumonia images, while the validation and test sets are more balanced.

Image: Examples of the dataset items (Source: Kaggle)

Experimental framework architecture: To decide on the most appropriate approach for us to the Active Learning Pipeline, we took inspiration from common approaches from literature. For example, Hao et al. (2021) suggested using a Convolutional Neural Network (CNN) as a classifier, fine-tuned on a small set of images prior to the active learning phase.

They have also suggested comparing the sampling strategy with a baseline model trained with random sampling. According to them, the goal should be to achieve high accuracy using a small percentage of the labeled dataset. Hemmer, Kühl, and Schöffer (2022) presented the idea of manipulating the last layer of a ready CNN architecture to achieve better results. Mahapatra et al. (2018) suggested the use of pre-trained model on a different dataset, and augmentation of the train set.

The architecture of the experimental framework we implemented is of a grid search that supports different sampling methods. At each experiment, the code loads the dataset, initializes the machine learning model, and runs an Active Learning loop. The loop iteratively queries unlabeled data, trains the model on the newly labeled data, and evaluates the model's performance. Results are stored and visualized for each model and sampling method.

Data Preprocess: In the data loading part, the code resizes images to 312×312 pixels, converts them to PyTorch tensors, and normalizes pixel values for better training. In addition, it splits the training dataset further into labeled and unlabeled sets, resulting in four data categories: train-labeled (the initial training set for the model), train-unlabeled (from which the model selects the informative samples to annotate), validation and test.

Core Loop: The core of the Active Learning loop receives the model and its configuration, labeled and unlabeled training sets, validation set, test set, and sampling strategy as input. It iterates for a specified number of iterations:

1. Starts with a copy of the model for each iteration (avoids accumulating previous training effects).
2. Trains the model on the current labeled train data and the labeled validation set.
3. Selects informative samples from the unlabeled pool based on the chosen strategy.
4. Updates the training data by adding the newly labeled samples and removes them from the unlabeled pool.
5. Evaluates the model on the test data and stores the metrics.

Finally, it returns a list containing the evaluation metrics for each iteration.

Sampling Methods: Our implementation encompasses the three sampling techniques outlined in the course:

1. Random Sampling: In each iteration, following the training of the model on the labeled dataset, the algorithm randomly chooses samples to be labeled and incorporated into the training set for the subsequent iteration.
2. Uncertainty Sampling: In each iteration, following the training of the model on the labeled dataset, the algorithm predicts the probability of each sample belonging to every class. The algorithm then selects those samples whose class association probabilities are closest to a 50-50 distribution.
3. Entropy Sampling: In each iteration, following the training of the model on the labeled dataset, the algorithm computes the entropy of the class probability distribution for each sample. It then identifies and selects the samples with the highest entropy values for labeling, as higher entropy signifies greater uncertainty or ambiguity, making these samples particularly valuable for labeling.

Additionally, we have developed a novel approach called "PCA then K-Means". This method selects samples according to the following procedure:

- For each sample, extract the weights from the final layer of the machine learning model prior to clustering (resulting in a weight vector for each sample).
- Apply Principal Component Analysis (PCA) to the weight vectors, reducing their dimensionality to a relatively low number (e.g., 3 or 10).

- Execute K-Means clustering to partition the unlabeled samples into K clusters, where K corresponds to the number of samples designated for labeling in each iteration.
- From each cluster, select one sample for labeling.

The fundamental premise of this approach is that distinct clusters encapsulate samples with varying characteristics, thus selecting images with diverse attributes from the unlabeled training set for labeling is deemed advantageous.

Although we initially conceived this idea independently, we subsequently discovered that there are related works that involve close collaboration, including those by Sener and Savarese (2017) and Kim and Shin (2022). Nevertheless, we have not identified any research that proposes a diversity-based methodology for sample selection in an Active Learning context that aligns with our proposed approach.

Models and Training: For the classification task, we employed the convolutional neural network architecture ResNet18, introduced by He, Zhang, Ren and Sun in 2016, with an adjustment made at the networks' final layer to accommodate a binary classification task. The training procedure is done in epochs, including several steps at each epoch: going over the data in mini-batches, calculating loss for each mini-batch, updating the model parameters, and finally printing the average loss for the epoch. After each epoch, it also evaluates the model's performance on both the validation and training sets, reporting the accuracy.

We sought to incorporate image augmentation to improve the model's robustness and overall performance; however, this strategy proved to be impractical due to the excessive time required for training, and it did not result in any performance enhancements, leading us to discontinue this approach. Additionally, we considered incorporating VGG16 (introduced by Simonyan and Zisserman in 2014) alongside ResNet18; however, the high computational cost associated with VGG16 led us to exclude it from our analysis, leaving it as a potential avenue for future research. The complete implementation for both ResNet18 and VGG16 are accessible and ready to run in our code repository.

Experiments

To examine the effectiveness of our Active Learning approach in the context of X-ray imaging classification versus traditional machine learning techniques, and our "PCA then K-Means" genuine approach in comparison with the random, uncertainty, and entropy approaches, we conducted several experiments.

Every experiment is defined by a set of parameters, we name as "the experiment configuration". The configuration includes, among other things, the machine learning model used, the number of samples used in the training phase, the number of samples to be labeled in every iteration, the number of Active Learning iterations, the number of epochs in every iteration, and the sampling method used in the Active Learning pipeline. Every such experiment results in accuracy, precision, recall, F1-score, ROC-AUC score, and confusion matrices.

Initially, we aimed to investigate the feasibility of utilizing Active Learning, specifically whether it is possible to attain results comparable to those achieved by training a model on the entirety of the dataset, even when using only a relatively small subset of the data. Consequently, the first baseline that was evaluated involves the conventional training of a machine learning model on the complete dataset, allowing us to ascertain the maximum potential accuracy metrics that can be achieved. To this end, we conducted training of the ResNet18 model using the full training dataset, subsequently gathering the corresponding accuracy metrics.

The evaluation of the test set revealed that the model produced the following outcomes (refer to Appendix A for comprehensive results):

Number of Epochs	Learning Rate	Test Accuracy	Test F1	Test Recall	Test Precision	Test ROC-AUC
3	1e-5	89.1%	0.919	0.905	0.912	0.942
5	1e-4	84.2%	0.800	0.997	0.888	0.937
5	1e-5	83.6%	0.797	0.989	0.883	0.925
3	1e-4	75.3%	0.717	0.997	0.834	0.865

Table 1: Test Set Metrics for ResNet18 with different configurations, Conventional Training Over the Entire Training Set

The model often demonstrated proficiency in identifying X-rays indicative of Pneumonia; however, it encountered difficulties in differentiating between labels for X-rays of healthy patients. This challenge can be attributed to the imbalance present in the dataset, which contains a greater number of images depicting patients with Pneumonia.

To evaluate the four sampling methods previously discussed, we conducted an experiment comprising four distinct versions, each differing solely in the sampling method specified in the experimental configuration. Each version commenced with an initial set of 100 observations, and over the course of 10 iterations, we incrementally added 90 observations, ultimately achieving a total of 1000 observations (refer to Appendix B for comprehensive results).

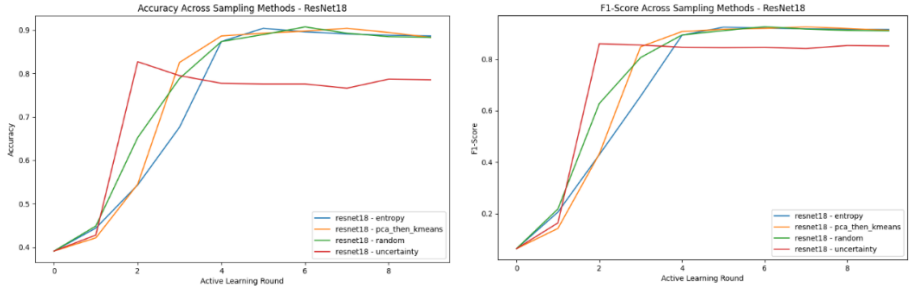

Chart 2: Test Set Accuracy and F1-Score across sampling methods, "100+90*10" experiments

Our investigation indicates that the Random, Entropy-based, and PCA then K-Means sampling techniques consistently surpass uncertainty sampling in terms of Accuracy, F1-Score, and Precision. This underscores their efficacy in selecting informative samples for active learning. In contrast, uncertainty-based sampling presents variable outcomes, excelling in Recall while falling short in Accuracy, F1-Score, and Precision, and achieving similar ROC-AUC scores to the other methods. It is evident that regardless of the sampling method employed, the algorithm effectively classifies Pneumonia cases; however, it encounters difficulties in accurately classifying X-ray images of healthy patients.

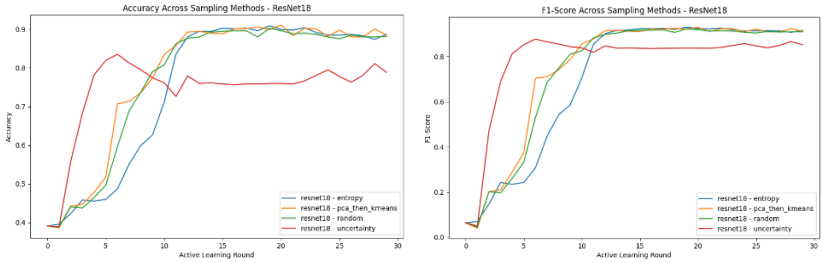

Chart 3: Test Set Accuracy and F1-Score across sampling methods, "100+30*30" experiments

In addition, we conducted another experiment comprising the four distinct versions, with an initial set of 100 observations, and over the course of 30 iterations, we incrementally added 30 observations, ultimately achieving a total of 1000 observations (refer to Appendix C for comprehensive results). The results suggest that the models converge approximately

to the same metrics values, but about 3 times slower, in correspondence with the decrease in the number of samples added to the train set per iteration. This insight suggests that the same labeling effort is needed to achieve similar results, regardless of the sampling method and the distribution of newly selected samples in each iteration.

Discussion

The results of our investigation into the enhancement of X-Ray imaging analysis through Active Learning (AL) provide valuable insights regarding the effectiveness of different sampling techniques in boosting model performance while reducing the need for extensive annotation. Our experiments indicated that employing Active Learning can achieve accuracy levels like those obtained from training on the complete dataset, even when working with a significantly smaller labeled data subset. This finding is particularly pertinent in the realm of medical imaging, where the scarcity of labeled datasets often arises from the considerable costs and time required for expert annotations.

A notable outcome of our study is the demonstrated efficacy of the "PCA then K-Means" sampling strategy. This method not only matched the performance of conventional techniques such as random sampling and uncertainty sampling in terms of accuracy, F1-score, and precision but also underscored the necessity of selecting a diverse array of samples from the unlabeled dataset. By clustering the samples and choosing representatives from each cluster, we ensured that the model encountered a wider variety of characteristics, which is essential for enhancing its generalization abilities.

Our findings suggest that although uncertainty sampling can achieve high recall rates, it frequently lacks in terms of accuracy and precision. This indicates that depending exclusively on uncertainty may be inadequate for tasks that require precise class differentiation, such as pneumonia detection. Consequently, future research should investigate hybrid methodologies that leverage the advantages of various sampling techniques to improve overall model efficacy .

In terms of potential advancements for subsequent studies, we propose examining the incorporation of generative models to create new, informative images for labeling purposes. This approach could further reduce the annotation workload by supplying additional training data that captures the intricacies of real-world situations. Another promising avenue for future investigation is the enhancement of the Active Learning loop itself. By substituting the ResNet18 model, which is relatively lightweight with a limited number of parameters, with more sophisticated models, it may be possible to achieve superior performance. The implementation of VGG16 is already included in our codebase, but it was not tested due to the impractical runtime constraints imposed by our computational resources.

In summary, our project successfully demonstrated the potential of Active Learning in the domain of pediatric radiology, specifically for the classification of chest X-rays into "sick" or "not sick." The findings underscore the importance of selecting informative samples and highlight the advantages of diverse sampling strategies. As we move forward, the integration of advanced techniques and a focus on optimizing the Active Learning process will be essential in further enhancing the effectiveness of machine learning models in medical imaging. By continuing to refine these methodologies, we can contribute to more efficient diagnostic processes and ultimately improve patient outcomes in the healthcare sector.

References

Chest X-Ray images (Pneumonia). (2018, March 24). Kaggle. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

Hao, H., Didari, S., Woo, J. O., Moon, H., & Bangert, P. (2021). Highly efficient representation and active learning framework for imbalanced data and its application to covid-19 x-ray classification.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
Hemmer, P., Kühl, N., & Schöffer, J. (2022). Deal: Deep evidential active learning for image classification. *Deep Learning Applications, Volume 3*, 171-192.

Kim, Y., & Shin, B. (2022, August). In defense of core-set: A density-aware core-set selection for active learning. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 804-812).

Mahapatra, D., Bozorgtabar, B., Thiran, J. P., & Reyes, M. (2018, September). Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 580-588). Cham: Springer International Publishing.

Nguyen, C., Huynh, M. T., Tran, M. Q., Nguyen, N. H., Jain, M., Vo, T. D., ... & Truong, S. Q. H. (2021, August). GOAL: gist-set online active learning for efficient chest X-ray image annotation. In Medical Imaging with Deep Learning (pp. 545-553). PMLR.

Sener, O., & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489.

Settles, B. (2009). Active learning literature survey.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wikipedia contributors. (2024, September 18). *Pneumonia*. Wikipedia. https://en.wikipedia.org/wiki/Pneumonia

X-ray Atlas: Chest X-ray | GLOWM. (n.d.). https://www.glowm.com/atlas-page/atlasid/chestXray.html

Appendix A – Link for full results of the "baseline" experiments

https://github.com/RoniFridman1/lab_094295/tree/main/outputs/baselines

Appendix B – Link for full results of the "100 + 90*10" experiments

https://github.com/RoniFridman1/lab_094295/tree/main/outputs/resnet18%20-%2010%20iter%20x%2090%20per%20iter

Appendix C – Link for full results of the "100 + 30*30" experiments

https://github.com/RoniFridman1/lab_094295/tree/main/outputs/resnet18%20-%2030%20iter%20x%2030%20per%20iter