Semi-Supervised Learning for Reducing the Annotation Load for Medical Imaging Datasets

Introduction:

Deep learning algorithms are used for a diverse set of tasks in the medical imaging domain. However, a majority of deep learning algorithms require a large amount of annotated datum to train models. Medical Imaging datasets are expensive to annotate and it is desired to develop methods to reduce annotation burden on radiologists. We want to use semi-supervised learning approaches to select a subset of samples from our existing datasets and improve the performance of DL models while reducing the amount of annotation burden.

Objective:

The primary objective of this project is to develop a semi-supervised learning framework that uses our one-of-its kind dataset in the histopathology domain and helps us in improving the diversity of the samples captured in this dataset. By leveraging this subset of labeled images along with the remaining unlabeled images, we aim to enhance the overall label quality and improve the performance of models that are trained on this dataset so that they can perform better in real-world settings.

Methodology:

The proposed methodology consists of the following steps:

1. Preprocessing: Prepare both the labeled and unlabeled images from the medical imaging dataset for modeling.
2. Model Deployment: Train models on the labeled samples and deploy these models onsite.
3. Scoring of model performance on the samples: Use the trained model to understand the certainty/quality of predictions for unlabeled samples. Using semi-supervised learning, get a subset of unlabeled samples that need to be annotated by the clinician. Get these subset of samples annotated by the clinician.
4. Model Finetuning: Finetune your model with the new labeled samples and compare the performance with the baseline model and/or models with a large number of labeled samples and small number of labeled samples from the semi-supervised paradigm.
5. Repeat steps 2 to 4 as and when you get new data and your model performs poorly.

Significance and Expected Outcomes:

- Improved Label Accuracy: The semi-supervised learning approach is expected to enhance the performance of the models by leveraging the labeled subset and propagating the labels to the remaining unlabeled images.
- Enhanced Research and Analysis: More accurate labels will enable researchers and medical professionals to perform reliable and meaningful analysis on medical images, leading to improved understanding and diagnosis of medical conditions.

- Application to other Datasets: The proposed will be a general framework that can be used generally for improving model performance in clinical settings with different medical imaging tasks.
- Facilitating Medical Decision-Making: Improved label accuracy can aid healthcare providers in making more informed decisions, leading to improved patient care and outcomes.

Project Steps:

The project will be conducted over a period of 6 months, and the timeline will be as follows:

- ➔ Literature Review. Dataset preparation and preprocessing.
- ➔ Model Deployment.
- ➔ Implementing semi-supervised learning algorithms.
- ➔ Experiments, evaluation, tweaks and fine-tuning.
- ➔ Finalizing the project report and preparing for presentation.

Conclusion:

This project proposal outlines a semi-supervised learning approach to improve the label accuracy in the MIMIC CXR dataset, assuming the NLP-based labels are correct. By leveraging a subset of images with assumed correct labels, we aim to propagate these labels to the remaining unlabeled images and enhance the overall label quality. The project's outcomes include improved label accuracy, insights into label quality, and validation of the proposed framework. The significance of this project lies in its potential to facilitate research, advance machine learning applications, and support medical decision-making in the analysis of chest X-ray images.

References:

- Johnson, Alistair EW, et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." Scientific data 6.1 (2019): 317.
- Johnson, Alistair EW, et al. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs." arXiv preprint arXiv:1901.07042 (2019).
- Van Engelen, Jesper E., and Holger H. Hoos. "A survey on semi-supervised learning." Machine learning 109.2 (2020): 373-440.