CS577 Intermediate Project Report: Multi-Label Medical Image Classification with MetaTeacher

Sukhmani Sandhu
A20513431
ssandhu3@hawk.iit.edu
Illinois Institute of Technology

Arpita Jadhav
A20523353
ajadhav11@hawk.iit.edu
Illinois Institute of Technology

Saurabh Rajput
A20523129
srajput2@hawk.iit.edu
Illinois Institute of Technology

***Abstract*—This is an intermediate report for our project in CS577 Deep Learning, Fall 2023. We are pulling bits of information from our project proposal, the original reference [1] and the other references, to give more detailed explanations. This is a report that documents our learning curve. It gives an insight on our journey in implementation so far.**

I. Introduction

A. Research Objectives and Synopsis

To build on what we have described in our proposal: Image Analysis in Healthcare continues to require progressive strides due to the challenge of large data-sets not being readily available because hand-annotation is expensive, time consuming and requires clinical proficiency– without which a primary issue is over-fitting. Another issue that we can derive is that of class imbalance. Suppose the data-set we work with has a large portion of heavily diseased patients with their respective pathological conditions or even a large portion of normal scans– it remains to be discussed how this might affect the performance of a model. This is a patent foundation of the problem that this paper seeks to solve by applying methods in transfer learning with a Teacher-Student approach. Furthermore, diagnosis requires accurate and precise predictions from models that are not infected with bias. For this purpose, the original paper- like many other recent papers in medical imaging use large data-sets (or combine moderate sized data-sets into a robust implementation)

Our study of the original work aims to implement the mathematical models of it on a small scale for the purpose of applying our knowledge of Deep Learning and building our understanding of the current research domains. Our implementation at this intermediate stage does not use all the data-sets mentioned in the reference paper– since we are trying to assess and tweak parameters based on classroom discussions within the course.

A key theme throughout the paper- we learn- is **domain adaptation**. Domain variability, in simple terms, is the difference between data collected from different sources. One of the reasons for this is a variation in the devices that collect the data since there is no one protocol to do so. Every MRI machine, for example, will capture different data of the same entity(patient) depending on treatment convention adopted, magnet specification and capacity. This can create a discrepancy when a model has to learn from data and apply it to new unseen data- in potentially a different environment. The reference paper compares their bench-marking with other available solutions. CNN's are a major stride in the field of medical image analysis but still do not perform with high accuracy on new examples. We struggle with accuracy being low despite there being abundant data collection efforts from a multitude of sources.

B. Considerations for Implementation

The paper takes into account these limitations to propose a novel framework for the context of multi-label image classification in healthcare- Semi-supervised Multi-source-free Domain Adaptation (SMDA). The paper defines a set of conditions/assumptions that the work is set in:

- a. Multiple source domain models are trained on multi-label image data-sets
- b. The source domain data is not available for adaptation
- c. The target domain has a limited number of labeled examples and a larger portion of unlabeled data

Note on (a): In existing research work, domain adaptation tasks currently require access to source domain data. They also majorly use a single source. The direct solution for this is the employment of an ensemble situation for the creation of a feature space which learns commonalities in the source data– this again, points to the need for knowledge of source domain data. There are other solutions named in the paper such as conditional adversarial discriminator networks which again, are not source-independent. Additionally, work in source-free domain adaptation assumes single-source which in turn creates bias. This is an aspect that cannot be ignored since hospitals have different demographics to pull data from. This cannot be applied to a general population.

C. Synopsis of Methodology

Therefore, in the pursuit to address all of these issues, the paper introduces the concept of transfer learning, specifically knowledge distillation from multi-source models to a target. That brings about a multi-teacher, one student framework.

Further, the paper optimizes using bi-level optimization to update the teachers and the student because models differ in reliability and the direction of update requires optimization as well.

In the MetaTeacher framework, each teacher model is trained on a particular labeled source. Then the student model is assigned by a random teacher. The approach now calls for the use of coordinated weight learning to weigh out the effect of each teacher for each target point.

II. Problem Statement

(as in reference paper) D_T is the vector(dimension m) for target= $(X_L^t, Y_L^t), X_U^t$ where Y_L^t is the set of label annotations for a small part of the target domain samples X_L^t. Here, X_U^t is the target domain samples that do not have annotations.

D_{S_i} is an iterable (let us say, i^th source domain sample)= (X_L^i, Y_L^i) where Y_L^i is the label for this i^th example X_L^i.

For the "SMDA" to be implemented, as proposed in the paper- when pre-trained source classifiers are applied to the target domain, the source data-set D_S will not be accessible for i= [0,n]

Now given the source classifiers and the target data D_T, the objective is defined to find a mapping in the target domain such that f: $X_U^t \rightarrow Y_U^t$ (predicted classes for target domain points X_U^t)

III. Methodologies Used

A. Coordinating Weight Learning

Above, we mention that MetaTeacher trains teacher models on different source domain data. Due to the vast difference in the distributions, they will present varying traits. This would lead to the classification probability of each teacher being inconsistent. So the paper proposes a way to find the weight contributions of each of the teachers to the final classification outputs.

Part (a) of the Figure 1 shows an overview of coordinating weight learning- where labeled target data is input to the student network and the output is B, from a feature extraction network g. The parameters that are needed for this mathematical model are c, number of channels; h-w, height and width. Then a maximum pooling is performed on B to get U which will retain the most essential information collected from each channel. The mapping also has two learnable variables- say u and v. This is put together and normalized to obtain a weight matrix W.

The loss used to initialize the student network L_W can be broken down into two parts- one term uses binary cross entropy loss and the other uses Kullback-Leibler divergence loss (for a measure of the distance in distributions between a fused student and teacher prediction)

B. Bi-level Optimization

First introduced in game theory, the concept of bi-level optimizations is divided into two tasks: upper and lower level optimization. Upper level optimization is constrained by lower level optimization. Here, the student is the former- it gives feedback to the latter (teachers) in terms of performance on the labeled data and the weight map we derive from above.

In this technique, a loss function that is a function of the parameters of the teacher network is defined. The respective updation of the teacher parameters belongs to the lower level optimization task and that of the student is upper level objective.

Essentially, since there is a mutual constraint on both updations that does not allow for one parameter set to start updating without the other set reaching its optimum (student)- we cannot employ simple gradient descent.

The solution proposed is- meta-learning. This involves approximating the problem into one step. In obtaining the pseudo-label we also binarize with the rule that if y¡0.5=0, else 1. These steps are shown in the algorithm in Figure 2

IV. Description of Data used in project

A. Relevance and Background of the data we chose

Chest radiography/x-rays are a frequent and cost-effective imaging modality used to analyze and diagnose a large number of health conditions (acute/chronic pulmonary, cardiac, thorax illnesses) owing to the large area this covers. This makes the need for accurate assessment in this subset of scans extremely significant. Another point we discovered is that the number of radiologists in the United States is decreasing, in ratio to the number of physicians. This further necessitates the application of deep learning models to aid in timely delivery of diagnostic acumen. Now, if we think of resource-limited areas, the predicament is worse. To quote a statistic directly: In 2015, a sparse 11 radiologists were available to Rwanda with a population of 12 million [3] and Liberia, with a population of 4 million, only had 2 radiologists [4].

B. Data Used by MetaTeacher Research

The reference paper implements their framework using five chest X ray data-sets that are accessible to the public [1]. However, about 2 of them are disclosed only to credentialed users and require signing a disclosure agreement in order to download and use due to the sensitivity of the data corresponding to real patient radiographic studies. The data-sets are namely: NIH-CXR14 (NIH Clinical Center, 14 disease labels), MIMIC-CXR (Beth Israel Deaconess Medical Center, text reports on the scans, ¡=1 studies), CheXpert (Stanford hospital, 65,240 patients, chest x-ray examinations), Open-i (Indiana University Hospital, stripped version, 3955 frontal images), Google-Health-CXR (manual annotation, derives data from NIH-CXR, 4000 notes)

C. Our current data and processing details

Up to this intermediate stage, we are working on a reduced implementation that tests our own application of the skills gained through class. So, to begin, we procured the NIH-CXR14 data-set. It consists of approximately 112,000 de-identified chest x-rays in PNG format. A few samples have been provided, for visualization purposes in Figure 3. These are front view images of 32,717 patients collected, as we have mentioned, from the NIH Clinical Center and are openly accessible. Also, we have provided another image of the co-occurrence statistics of the 14 disease categories, as provided with the data-set [6]

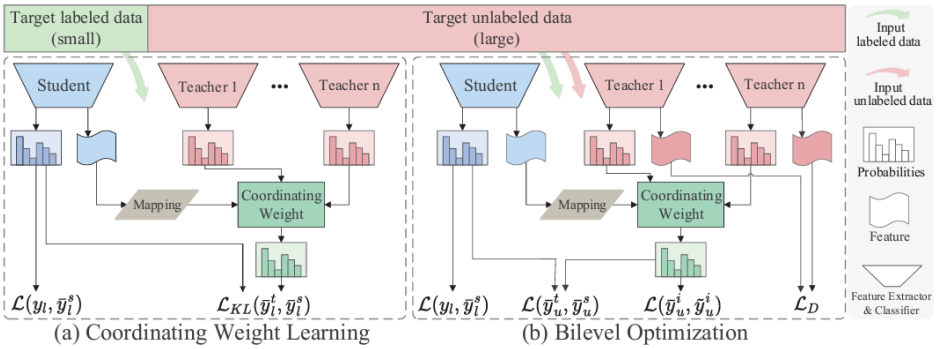


Fig. 1. Learning the coordinating weight mapping which will be used for updating the teachers. (b)Alternately updating the teacher and student models.

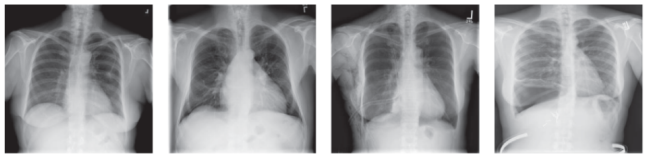
Algorithm 1 Our proposed MetaTeacher method.

Require: Student network parameters $S^{(0)}$, teacher network parameters $T_1^{(0)} \sim T_n^{(0)}$, labeled data (x_l^t, y_l), unlabeled data (x_u^t), hyperparameters α, β, γ, mapping updating interval $\mathcal{T}$.
Ensure: Optimized student model $S^{(N)}$.

function METATEACHER($S^{(0)}, T_1^{(0)} \sim T_n^{(0)}, \alpha, \beta, \gamma, \mathcal{T}$)
  $S^{(0)}$, mapping $\leftarrow$ *Coordinating Weight Learning*
  for $t = 0 \rightarrow N-1$ **do**
    Upper-level optimization:
    Compute gradient $\nabla_{\theta_{S^{(t)}}} \mathcal{L}_u$
    Update the student: $\theta_{S^{(t+1)}} \leftarrow \theta_{S^{(t)}} - \eta_S \nabla_{\theta_{S^t}} \mathcal{L}_u$ ▷ Eq.(8)
    Lower-level optimization:
    Compute gradient $\nabla_{\theta_{S^{(t+1)}}} \mathcal{L}_l$
    for all $T_1^{(t)} \sim T_n^{(t)}$ **do**
      Compute gradient $\nabla_{\theta_{T_i^{(t)}}} \mathcal{L}\left(\tilde{y}_u^i, \tilde{y}_u^i\right)$
      Compute gradient $\nabla_{\theta_{T_i^{(t)}}} \mathcal{L}_D$ ▷ Eq.(10)
      Update the i-th teacher: $\theta_{T_i^{(t+1)}} \leftarrow \theta_{T_i^{(t)}} - \eta_{T_i} \cdot$ ▷ Eq.(11)
        $\left([(\nabla_{\theta_{S^{(t+1)}}} \Gamma_l)^T \cdot \nabla_{\theta_{S^{(t)}}} \Gamma_u]^T \cdot \nabla_{\theta_{T_i^{(t)}}} \mathcal{L}\left(\tilde{y}_u^i, \tilde{y}_u^i\right) + \gamma \nabla_{\theta_{T_i^{(t)}}} \mathcal{L}_D\right)$
    end for
    if $t \% \mathcal{T} = 0$ **then**
      mapping $\leftarrow$ *Coordinating Weight Learning*
    end if
  end for
  return $S^{(N)}$
end function

Fig. 2. Main Algorithm

From: Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification



(a) "No Finding" **(b)** "Cardiomegaly" **(c)** "Pneumothorax" **(d)** "Pneumothorax"

Fig. 3. 4 samples from the NIH-CXR14. All images are labeled with a maximum of 14 pathological diagnoses or "No Finding" [6]

Our pre-processing will follow the same implementation as in the reference paper. To create uniformity, images from all data-sets are scaled to size 128x128. Horizontal flipping and random cropping are some of the data augmentation methods employed to increase the training data-set size.

V. Work Done so far

In the paper, four transfer scenarios are defined as: NIH-CXR14, CheXpert, MIMIC-CXR to Open-i; NIH-CXR14, CheXpert, MIMIC-CXR to Google-Health-CXR; CheXpert, MIMIC-CXR to NIH-CXR14; NIH-CXR14, CheXpert to Open-i. This is formatted as *from* **teacher1,teacher2. . . teacherN** *to* **student**.

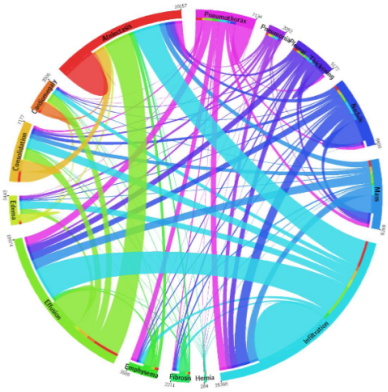


Fig. 4. Distributions of the 14 labels, co-occurrence

Now, for this stage of our implementation in the class project, we have defined a slightly simpler transfer scenario, with the data we have procured, due to not having access to the other data without credentials. We split the NIH-CXR14 into two parts (for two teachers) and the Open-i set as the student. **ie. NIHCXR14(a), NIHCXR14(b) to Open-i**

Note: that we also decided to use a single scenario and limit our datasets since we seek to explore the implementation from a simple, bottom to top approach– to facilitate learning.

The original code implementation has several models, we start with creating our own student model, basecnn.py to replace the existing myNet.py in the student model. For the teacher model we have implemented the alexnet.py and googlenet.py models as provided in the original code by the MetaTeacher team.

For optimizing, we switched out of using SGD and instead used Adam optimizer: For the student model, we use an initial learning rate of 0.01. We use 0.01 for the teacher sets.

Our goal with reducing our compute load is to thoroughly understand the framework that was proposed by the team of researchers for MetaTeacher.

We also discovered that we require GPU resources and the problem with migrating to Google Colaboratory is memory volatility and wait-queues for GPU access.

VI. Work that remains

By the final stage of this project, we aim to have implemented all the aspects of the MetaTeacher approach. We would start with acquiring more data, to create a solid transfer scenario. In case we are able to procure enough data, we shall create enough scenarios to draw a bench-marked comparison on the hyper-parameters that we used.

The specific criterion for evaluating performance, used in the reference text is AU ROC (Area Under Receiver Operating Characteristic) curve.

Note: We are also trying to potentially acquire GPU access from college resources in order to truly test the implementation for the high accuracy that it reports.

VII. Team Contributions and Responsibilities

This section will be updated with every step we take into this project. At present, with respect to this deliverable:

- Sukhmani: Data Acquisition, Writing/Compiling Report, Student model creation
- Saurabh: Teacher Model creation, Code management/compilation
- Arpita: Data Pre-processing, Optimizer selection, Debugging

References

[1] Wang, Z., Ye, M., Zhu, X., Peng, L., Tian, L., Zhu, Y. (2022) MetaTeacher: Coordinating Multi-Model Domain Adaptation for Medical Image Classification.

[2] https://physionet.org/content/mimic-cxr/2.0.0/ MIMIC-CXR Database Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, Steven Horng

[3] David A Rosman, Jean Jacques Nshizirungu, Emmanuel Rudakemwa, Crispin Moshi, Jean de Dieu Tuyisenge,Etienne Uwimana, and Louise Kalisa. Imaging in the land of 1000 hills: Rwanda radiology country report. Journal of Global Radiology, 1(1):5, 2015

[4] Farah S Ali, Samantha G Harrington, Stephen B Kennedy, and Sarwat Hussain. Diagnostic radiology in Liberia: a country report. Journal of Global Radiology, 1(2):6, 2015

[5] https://www.nature.com/articles/s41598-019-42294-8 BALTRUSCHAT, I. M., NICKISCH, H., GRASS, M., KNOPP, T., AND SAALBACH, A. Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports 9, 1 (2019), 1–10.

[6] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly- Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471,2017