CS577 Final 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 final report for Group 21's term project in CS577 Deep Learning, Fall 2023. We are pulling bits of information from our intermediate report as well, the original reference [1] and the other sources, 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. The Github repository for our project's uploads and experiments is available at https://tinyurl.com/4pde5b52**

I. Introduction

A. Research Objectives

To build on what we have described in our intermediate report: Image Analysis in Healthcare continues to require progressive strides due to the challenge of staggering/poor performance on new unseen data from multiple sources and large datasets not being readily available because hand-annotation is expensive, time consuming and requires clinical proficiency– without which a primary issue is over-fitting. We note that it is not a dearth of data that is the problem- but the lack of enough data that has been annotated clinically. And collation and training of said data is a subsequent activity that is being optimized using efficient deep learning techniques. Another issue that we can derive is that of class imbalance. Suppose the dataset we work with has a large portion of heavily diseased patients with their respective pathologies or even a large portion of normal scans– it is worth discussing 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 datasets (or combine moderate sized datasets into a robust implementation)

Our term paper based on 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 through the course and building our understanding of the current research domains. Our implementation at this stage does not use all the datasets mentioned in the reference paper– since we are trying to assess and tweak parameters based on classroom discussions over the semester.

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.

The MetaTeacher framework introduces three components to its functioning: (1) Learn-able coordinating scheme for learning the weights, for adaptive source models, to provide different directions for updates in the multiple teachers, based on their contributions in weight (2) semi-supervised bi-level optimization strategy for the weight updation of the teachers and student, which is critical for solutions in the low-supervision challenge because it can help leverage collaborative learning of the source models, (3) Mutual feedback mechanism between the source models(teachers) and the target model(student) for coherent learning. Meta-learning, an underlying theme, involves training a model on multiple tasks to make it more adaptive to new tasks. In domain adaptation- this can help a model rapidly adapt to a new domain with very limited labeled data (from this target site.)

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: 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:

- Multiple source domain models are trained on multi-label image data-sets
- The source domain data is not available for adaptation

- The target domain has a limited number of labeled examples and a larger portion of unlabeled data

Note: 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.

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 bilevel 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.

B. Related Previous Work

While the work done by the MetaTeacher team claims to be a pioneering first effort in the direction of Multi-Source-Free and Semi Supervised Domain Adaptation in Transfer Learning, there are several different strides in the direction. The current models, which we will expound on in detail, lack in one or the other aspect. Briefly, Shallow and Deep Unsupervised Domain Adaptation approaches still need access to source domain data. Source-free DA uses only single source domains. Multi-source DA– approaches are shallow and still require access to source domain data.

Hence, for better domain alignment we have semi-supervised DA which assumes a small number of labeled samples in the target domain. Specifically, previous work is as described below:

- **Unsupervised Domain Adaptation for Medical Image Classification**
  - Shallow UDA Approaches:
    (1) Source Domain Instance Weighting: This method involves adapting two routes in the domain adaptation process [7]
    (2) Feature Transformation: Another approach within shallow UDA, it focuses on transforming features to enhance adaptation [8]
  - Deep UDA Approaches:
    (1)Domain Alignment Based: The primary strategy minimizes domain differences between source and target domains, currently a widely adopted method. Example: Gao et al. [9] utilized central moment difference matching for brain MRI data classification.
    (2) Pseudo-labeling Based: This strategy involves generating dummy data to retrain the target model. Example: Bermúdez Chacón et al. [10] employed normalized cross-correlation to generate soft labels for the target domain.
    (3) Domain Alignment with Multi-label Regularization Term: Specifically tailored for multi-label classification, this work incorporates domain alignment with a multi-label regularization term.
  - Considerations for Practical Scenarios: Challenge: Existing UDA methods typically focus on single-source domains.
    Practical Scenario: Addressing the common occurrence of multi-source domains in real-world applications.It is to be noted that all the mentioned UDA methods maintain the integrity of the source domain model without updates.

- **Source-Free Domain Adaptation in Medical Image Analysis**
  - Generative Approach:
    (1) Generation of Target-Style Training Samples: This approach, falling under source-free domain adaptation, focuses on generating target-style training samples to train the prediction model.
    Challenge: Learning to generate features is acknowledged as a tough aspect, limiting the scope of this approach.
  - Pseudo-label Approach:
    (1) Pseudo-label Generation through Source Domain Model: This alternative approach involves generating pseudo-labels using the source domain model, recognized for its simplicity and general applicability.
    Recent Success: The pseudo-label approach has demonstrated significant success in the machine learning community.
  - Application in Medical Image Analysis—Focus on Image Segmentation:
    (1) Mutual Information Maximization for Image Segmentation:
    Example: Bateson et al. [11] maximized mutual information between target images and their label predictions for spine, prostate, and cardiac segmentation.
    (2) Source-Free Domain Adaptive Image Segmentation: Example: Vibashan et al. [12] implemented source-free domain adaptive image segmentation by generating pseudo-labels and applied self-training methods for task-specific representation.
  - Consideration for Single-Source Domain Case:
    Observation: Existing works in source-free domain adaptation within medical image analysis primarily

focus on single-source domain scenarios.
Limitation: Limited research on multi-source-free domain adaptation; many existing works adopt the method of generating trusted pseudo-labels.

- **Semi-Supervised Domain Adaptation (SSDA)**
  - Connection: The reference paper described problem shares similarities with Semi-Supervised Domain Adaptation (SSDA), which operates under the assumption of having a small number of labeled samples in the target domain.
    Advantage: Unlike Unsupervised Domain Adaptation (UDA), the utilization of a few labeled samples in SSDA facilitates improved domain alignment.
    Challenges in Applying Classical Methods:
    Issue: Directly applying classical semi-supervised learning methods to SSDA, considering the shift in domain distribution, may result in sub-optimal performance.
  - Representative SSDA Approaches:
    (1) Subspace Learning:
    Some SSDA works are based on subspace learning, aiming to align domains effectively.
    (2) Entropy Minimization:
    Entropy minimization is employed in certain SSDA approaches to enhance performance.
    (3) Label Smoothing:
    Label smoothing is a technique utilized in SSDA to improve the robustness of the model.
    (4)Active Learning:
    Active learning strategies are employed in SSDA to intelligently select and label the most informative samples.
  - Distinguishing Feature of the reference paper's method:
    Unlike existing SSDA methods, MetaTeacher incorporates meta-learning.
    Feedback Signal: Our method utilizes the performance on labeled target data as a feedback signal, offering a unique perspective in addressing the domain adaptation challenge.
- **Multi-source Domain Adaptation for Medical Image Classification**
  - (1) Distribution Alignment:
    Description: This strategy in multi-source domain adaptation (MSDA) computes the statistical discrepancy between multiple source domains and the target domain.
    Implementation: Combines predictions from all source domains after aligning distributions.
  - (2) Adversarial Learning:
    Description: The second MSDA strategy involves training a domain discriminator and compelling the feature extraction network to learn domain-invariant features.
    Implementation: Aims to confuse the domain discriminator by forcing the extraction of features that are invariant across domains.
  - Shallow DA Models in Medical Image Classification:
    (1) Mapping to Common Latent Space:
    Example: Wang et al. [13] proposed a model that maps multiple source and target data to a common latent space for autism spectrum disorder classification.
    (2) Multi-domain Transfer Classifier:
    Example: Cheng et al. [14] developed a multi-domain transfer classifier for the early diagnosis of Alzheimer's disease.
  - Limitations of Existing Approaches:
    Requirement: Current MSDA strategies often require access to source domain data.Existing shallow DA models may not be suitable for solving the proposed problem of Single Multi-source Domain Adaptation (SMDA).
    Challenges in Multi-source Domain Adaptation:
    Observation: Current teacher-student domain adaptation methods in medical and machine learning communities predominantly address the single-source domain case.
    Extension Challenge: Extending these methods to multi-source domains poses a challenging multi-objective optimization problem.
- **Teacher-Student Domain Adaptation Models in Medical Image Analysis**
  - Focus: Teacher-student domain adaptation models are designed to address the Unsupervised Domain Adaptation (UDA) problem.
    Consistency Strategies: Typically, these models propose multiple consistencies to resolve UDA challenges.
    Limited Attention in Medical Image Analysis:
    Observation: Despite their prevalence in other domains, teacher-student based domain adaptation methods have received limited attention in the context of medical image analysis.
    Example Method of Teacher-Student Model in Medical Image Segmentation: Proposes a semi-supervised learning based UDA method for medical image segmentation.
    Consistency Loss Minimization: Focuses on minimizing the consistency loss between predicted results of the student and teacher models for unlabeled samples in the target domain.
    Implementation: Employs the mean-teacher scheme, updating the network through the exponential moving average of the student network weights.
    Demonstration: Experimentally performed on the SCAM (Spinal Cord Anatomy MR Image) dataset to showcase effectiveness.
  - Limitations and Unexplored Territory:
    Assumption: Existing teacher-student domain adap-

tation methods assume a single-source domain.
Gap in Research: As of now, there is no known work on multi-source teacher-student domain adaptation.
Feedback Signal Utilization: The mean-teacher approach is noted for not fully leveraging the feedback signal from the target domain, resulting in limited performance improvement.

II. Synopsis of Methods

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.

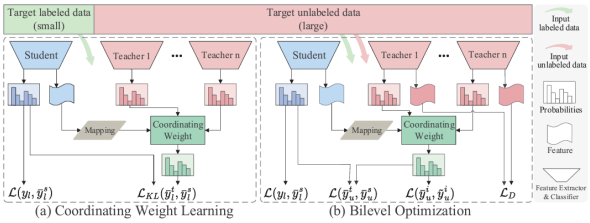


Fig. 1. Learning the coordinating weight mapping which will be used for updating the teachers. (b)Alternately updating the teacher and student models.

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 $y_i 0.5=0$, else 1. These steps are shown in the detailed algorithm below.

III. Summary of Results

The reference paper's methods achieve unparalleled performance to all state of the art systems and alternatives- on all transfer scenarios based on the five large, credential access chest x-ray datasets. The paper aims to leverage the knowledge of source models adaptively and aims to maximize their complementary benefits collectively for the challenge of limited supervision.

In our class project's basic and simple attempt at implementing the Metateacher framework, we achieved high accuracy on training the transfer scenario within the NIHCXR dataset. (In our transfer scenario we split the dataset into parts and distribute them across different teacher/student models for training and test on a smaller yet subset.)

IV. Problem Description

Framework Details: As illustrated in Fig.1, is built upon a multi-teacher and one-student architecture. Initially, multiple teacher models undergo pretraining specific to each source domain. Subsequently, the student model is initialized, randomly selecting one of the teacher models. Both teacher and student models consist of a feature extractor and a multi-label classifier. The classifier comprises a fully connected layer, receiving an expanded one-dimensional feature as input and producing label probabilities as output. The primary objective function centers on the error loss between the predicted output and the ground truth.

For ease of explanation: we will focus on the output losses, depicted in Figure 1. There are six quantities that we will describe in detail, below- and explain how they are mathematically acquired and used while walking through the algorithm.

A. Problem Statement

(as in reference paper) D_T is the vector(with 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)

V. Description of the Data

A. Relevance and Background of the Data

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, only one of them (NIHCXR) is truly available for public research and the rest require registration- are credentialed, need extended training and disclosure procedures for procurement 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)

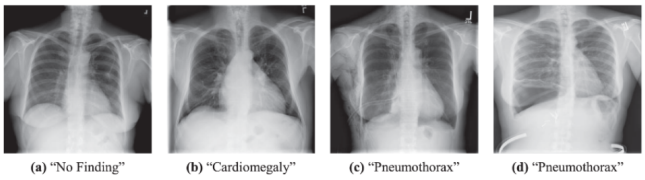
C. Current data and pre-processing details

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] The colored bands are basically a span showing each color as a co-occurence of one or more disease labels in the same patient. The numbers on the circumference are the instance of said overlaps.

Our transfer scenario is built as T1, T2 to S. For T1, we use data from 120 patients which is nearly 520 images. For T2, we use data from 120 patients which is nearly 460 images. For S, we use data from a 100 patients which is nearly 370 images. For the test folder, we use data from 50 patients which is nearly 200 images. We modify the CSV files to only include information the images we are using.

Our pre-processing of the images was minimal and included a very standard normalizing and did not involve flipping and scaling to augment the data. This remains to be worked with, in the future. Again, we are simply trying to understand the implementation provided in the paper.

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



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

Fig. 2. 4 samples from the NIH-CXR14. All images are labeled with a maximum of 14 pathological diagnoses or "No Finding" [6]

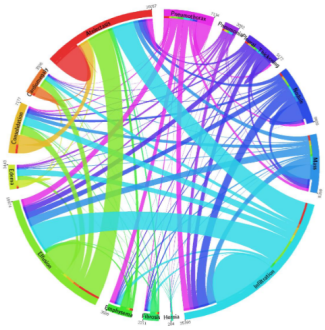


Fig. 3. Distributions of the 14 labels, co-occurrence

VI. Methodology: Algorithm Walk-through

A. Mathematical Derivations/ Formulations

For the algorithm to be implemented, some of the key components are the mathematical formulations, how they are computed and what they represent in our learning process.

If we refer to Figure 1, the model architecture, we see that in order to implement the three main components- coordinated

Algorithm

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: $\theta_{S^{(t+1)}} \leftarrow \theta_{S^{(t)}} - \eta_S \nabla_{\theta_{S^t}} \mathcal{L}_u$
    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 $\nabla_{\theta_{T_i^{(t)}}} \mathcal{L}\left(\overline{y}_u^i, \widetilde{y}_u^i\right)$
      Compute gradient $\nabla_{\theta_{T_i^{(t)}}} \mathcal{L}_D$
      Update the i-th teacher: $\theta_{T_i^{(t+1)}} \leftarrow \theta_{T_i^{(t)}} - \eta_{T_i} \cdot$
        $([(\nabla_{\theta_{S^{(t+1)}}} \Gamma_l)^T \cdot \nabla_{\theta_{S^{(t)}}} \Gamma_u]^T \cdot \nabla_{\theta_{T_i^{(t)}}} \mathcal{L}\left(\overline{y}_u^i, \widetilde{y}_u^i\right) + \gamma \nabla_{\theta_{T_i^{(t)}}} \mathcal{L}_D)$
    end for
    if $t \% \mathcal{T} = 0$ **then**
      mapping $\leftarrow$ *Coordinating Weight Learning*
    end if
  end for
  return $S^{(N)}$
end function

Fig. 4. Main Algorithm

weight learning, feedback and bi-level optimization we obtain 6 different loss metrics.

In order, from left to right- we have:
$\mathcal{L}(y_l, \overline{y}_l^s)$, $\mathcal{L}_{KL}$, $\mathcal{L}(y_l, \overline{y}_l^s)$, $\mathcal{L}(\overline{y}_u^t, \overline{y}_u^s)$, $\mathcal{L}(\overline{y}_u^i, \widetilde{y}_u^i)$ and $\mathcal{L}_D$. For your information, $\overline{y}$ represents the predicted labels.

In Coordinated Weight Learning:
We know that the goal of this process is to ensure that we handle inconsistencies in probabilities of the teacher models based on the small subset of target domain data by computing different directions for optimizing each teacher. This is done by assessing the weight contributions of individual teachers. We do this by generating a weight matrix with the few labeled instances of the target domain data.

Therefore, when we input the labeled samples into the student network (x_t^l)), we get B= g(x_t^l)) where g is the feature extractor. In order to retain the most significant information from B, we perform a a maximum pooling on this feature map to eventually obtain ψ.

A little more mathematical manipulation is required in order to output a normalized weight matrix W:

$$W_{j,k} = \frac{exp(\phi_{j,k})}{\sum_{z=1}^{n} exp(\phi_{z,k})}.$$

Now we also obtain something called the fused prediction P which is a prediction on the labeled data we are inputting. This is done by taking the predictions of all the teachers into a matrix and performing a Hadamard Product between P and W. This gives us: $\overline{y}_t^l$

We denote the student model's prediction on this sample as $\overline{y}_s^l = f_S(x_t^l)$; $_S$) and obtain the equation for the loss below:

$$\mathcal{L}_W = \mathcal{L}\left(\overline{y}_l^s, y_l\right) + \alpha \mathcal{L}_{KL}\left(\overline{y}_l^t, \overline{y}_l^s\right) + \beta\left(\|\mu\| + \|\nu\|\right)$$

If we observe closely, this is summing up two of the losses from above: the binary cross entropy loss and the KL loss, Kullback-Leibler divergence loss- measuring the difference in the fused teacher and student prediction distributions. We also use some parameters for balancing.

Now, from Bi-Level Optimization:
We know that this consists of an (A) Lower Level Optimization task and (B) Upper Level Optimization task. **A** is a constraint on **B**, that is,updating the parameters of Teachers and the Student is the task objective. Here, B[student] provides feedback signals to A[teachers] through performance on the weight mapping and labeled data.

So here, we define a loss function based on pseudo-labels based on weight mappings as $\mathcal{L}(\overline{y}_u^t, \overline{y}_u^s)$, and a similar loss is defined for the labeled target sample set as $\mathcal{L}(y_l, \overline{y}_l^s)$ based on teachers' and student's θ.

The lower level objective function is then formulated but we cannot proceed with optimization with simple gradient descent because we cannot update teacher parameters unless we have completely optimized the student weight parameters. So, a one-step approximation is computed and a a final objective function is obtained:

$$\Gamma_l\left(\theta_{T_1}, \cdots, \theta_{T_n}, \theta_S - \eta_S \cdot \nabla_{\theta_s} \Gamma_u\left(\theta_{T_1}, \theta_{T_2}, \cdots, \theta_{T_n}, \theta_S\right)\right).$$

We then derive updation rules with $\mathcal{L}(\overline{y}_u^i, \widetilde{y}_u^i)$ where $\widetilde{y}_u^i$ is the pseudo-label is acquired with a binarization of the predictions ie. 0 when less than 0.5 and 1 for all other cases.

Furthermore, pointing to the next loss that is output from the bilevel optimization in the model architecture, $\mathcal{L}_D$ is the divergence loss that is used to prevent the teachers from all optimizing in the same direction. So we have to ensure that the predictions of the teachers are as distant as practically allowed: here L2 is the norm applied to B, which is the pooled feature

$$\mathcal{L}_D = -ln \sum_{j=1, j\neq i}^{n} \mathcal{L}_2\left(B_{T_i}\left(x_u^t; \theta_{T_i}\right), B_{T_j}\left(x_u^t; \theta_{T_j}\right)\right)$$

map from the multiple teachers. The final rule is acquired as below:

$$\theta'_{T_i} = \theta_{T_i} - \eta_{T_i} \cdot \left(\left[\left(\nabla_{\theta_{S'}} \Gamma_l\right)^T \cdot \nabla_{\theta_S} \Gamma_u\right]^T \cdot \nabla_{\theta_{T_i}} \mathcal{L}\left(\tilde{y}_u^i, \tilde{y}_u^i\right) + \gamma \nabla_{\theta_{T_i}} \mathcal{L}_D\right)$$

B. Iteration Steps

We initialize the student network parameters and teacher network parameters. We receive labeled data for training and unlabeled data and decide on the hyper-parameters α, β, γ, τ

S(0) is used for coordinated weight learning and then we move into optimizing. For iterations t=0 to t=N-1, We perform upper level optimization by computing the gradient and updating the student. We then move into lower level optimization, compute the gradient between predicted and pseudo labels for the unlabeled data for all teachers along with divergence loss. Finally we update the i-th teacher using the Final Rule equation above. Lastly, If tτ=0, then update the mapping through coordinating weight learning and return the optimized student model $S^{(N)}$

VII. Results

A. State-of-the-Art VS. MetaTeacher

Since this paper is a seminal work, an experimental comparison to the problem setting does not exist. In order to observe any speed-up and performance benchmarks the paper chooses four categories of methods:

-Source Only: the direct application of one teacher model to the target data [which is something that is a part of our own implementation as well]

-Fine Tune (Average) where each teacher network is being fine tuned using the small sample of labeled target data- and then averaging on their predictions

-State-of-the-Art Multi Source Free domain adaptation techniques [DECISION, SHOT-best, CAiDA]

-Semi-supervised domain adaptation methods [MME, ECACL] where the paper operated with the assumption that they use the same labeled target domain data across the methods. We then take only the BEST result, not average- from all models, in this comparison case.

The table below shows the results on the transfer scenario from NIH-CXR14, CheXpert, MIMIC-CXR to Open-i where the metric being administered is the Area Under Receiver Operator Characteristic.

The source only(N),source only(M), source only(C) refer to the single source scenarios for each of the teacher data sets.

MetaTeacher(all) is the proposed final method. It is clearly performing outstandingly.

MetaTeacher(w/o mapping) is the proposed framework stripped of the coordinating weight learning, and the optimization is done using averages.

B. Our Results in Comparison

Due to the limitation of data access for our own work as a team in the Deep Learning Course, we used the Source Only metric, to set a comparison to the above framework.

These are our per epoch performances, along with the learning curves for one of our students model (using ResNet 18)

VIII. Conclusions and Future Scope

A. Discussions of Results Observations

When we analyze the performance of the MetaTeacher framework in comparison to the state-of-the-art methods, we observe that the performance is still competitive without coordinated weight learning (w/o mapping case).

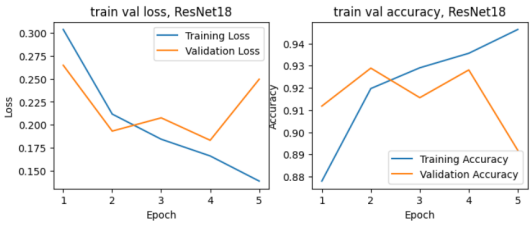
This is because for the student updates, predictions being averaged from the teacher models is necessarily beneficial for the student model's performance. Further, the fixed W because of no updating, is involved in the update of teachers- so bi-level optimization itself contributes more gain to the performance "en full" albeit coordinated weight learning will be able to classify which disease label is accurate, from which teacher. So, a particular teacher model may be exceptional at detecting/predicting one specific label. We note that the predictions in each category of disease is the best teacher's prediction, for instance, Pneumonia (from the above table)

Parameters, tuning and analysis: As mentioned earlier, the MetaTeacher algorithm involves the use of three balance parameters that can be tested on for their influence. α provides optimal performance and updation of teacher models at 0.5. If we go any lower, the weight matrix W is not trained well. This in turn will lead to the direction of optimization for teachers indeterminable. If β is made too large, the coordinating weight learning module would fail to capture the relationships between the different source domains and would render the module entirely unproductive. If instead, we use too small a value, say 0 we would face over-fitting and the weight learning

Method	Atelectasis	Cardiomegaly	Effusion	Consolidation	Edema	Pneumonia	*Average*
DECISION [1]	**83.27**	91.55	96.18	97.02	92.74	89.24	91.67
CAiDA [14]	82.45	92.16	95.12	95.92	89.89	90.37	90.99
SHOT-best [35]	81.48	91.22	94.19	95.10	88.96	89.58	90.09
MME [50]	82.44	90.82	95.46	96.07	90.26	87.20	90.38
ECACL [30]	82.60	92.18	96.32	95.97	90.70	89.61	91.23
Source Only(N)	83.09	87.20	96.11	95.10	86.87	77.40	87.63
Source Only(C)	82.26	87.64	94.71	96.61	90.22	75.12	87.76
Source Only(M)	80.63	91.31	94.87	94.53	84.91	82.78	88.05
Fine-tune(*average*)	82.14	88.71	95.32	95.52	88.77	78.48	88.16
MetaTeacher(w/o *mapping*)	79.99	**92.64**	**98.22**	93.64	**95.50**	84.54	90.76
MetaTeacher(w/o *update*)	81.98	90.72	95.76	95.51	89.40	82.53	89.32
MetaTeacher(*all*)	81.72	92.59	96.25	**97.64**	94.52	**94.33**	**92.84**

Fig. 5. comparisons to state-of-the-art

```
Epoch 1/5, Train Loss: 0.3038, Train Accuracy: 0.8780
Epoch 1/5, Validation Loss: 0.2648, Validation Accuracy: 0.9118
Epoch 2/5, Train Loss: 0.2117, Train Accuracy: 0.9197
Epoch 2/5, Validation Loss: 0.1931, Validation Accuracy: 0.9289
Epoch 3/5, Train Loss: 0.1843, Train Accuracy: 0.9291
Epoch 3/5, Validation Loss: 0.2076, Validation Accuracy: 0.9156
Epoch 4/5, Train Loss: 0.1661, Train Accuracy: 0.9356
Epoch 4/5, Validation Loss: 0.1832, Validation Accuracy: 0.9281
Epoch 5/5, Train Loss: 0.1388, Train Accuracy: 0.9463
Epoch 5/5, Validation Loss: 0.2496, Validation Accuracy: 0.8918
```



will work excellently in some cases but poor on more. γ specifically influences the divergence loss. When this is at 0.01 it is at peak performance, with gradual increment, performance starts to decline.

B. Contributions of the Project, and Team

This reference we selected for this term has presented us with a novel framework for the express advancement of medical image analysis using transfer learning, specifically multi-source-free, semi-supervised domain adaptation.
Each of us contributed in ways that helped us learn throughout the semester while applying our knowledge.

- Sukhmani: Data Acquisition, Writing/Compiling Reports, Student model creation, Hyperparameter Analysis
- Saurabh: Teacher Model creation, Repository management
- Arpita: Data Pre-processing, Optimizer selection, Debugging

C. Lessons We Learnt

When we were introduced to transfer learning, we started exploring knowledge distillation and teacher-student models. Specifically, we read through the research suggested in class presentations. We then took it upon ourselves to understand the core principles and advancement that we can achieve in a small data, CPU/GPU setting.

With a fairly naive understanding of the constituents of transfer learning (deep learning) we were able to build some foundational familiarity with this piece of theoretical work that is novel and striving to outperform.

We worked on a simple and rustic, unadorned model building- from bottom to top. We implemented the basic methodologies to create the transfer scenarios specified in the reference paper.

D. Potential Scope of Future Work

We are looking to continue our work to build on this model's framework and delve more deeply into teacher-student domain adaptation in the realm of medical image analysis, as part of dedicated independent research. We are yet to explore the use of stronger system GPUs with high capacity. Access to more datasets is something to look for: acquisition is also possible- with interdisciplinary involvement of the Biomedical Engineering Department at Illinois Institute of Technology.

Tackling Semi-Supervised Domain Adaptation (SSDA) revealed the delicate balance between leveraging a small number of labeled samples in the target domain and the challenges posed by domain distribution shifts. Navigating this balance illuminated avenues for future research and algorithm refinement.Our journey highlighted the effectiveness of incorporating meta-learning and bi-level optimization techniques in domain adaptation frameworks. These additions significantly enhanced the adaptability of our model, providing a dynamic response to varying domain scenarios.

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

[7] VAN OPBROEK, A., VERNOOIJ, M. W., IKRAM, M. A., AND DE BRUIJNE, M. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. Medical image analysis 24, 1 (2015), 245–254.

[8] LI, W., ZHAO, Y., CHEN, X., XIAO, Y., AND QIN, Y. Detecting alzheimer's disease on small dataset: A knowledge transfer perspective. IEEE journal of biomedical and health informatics 23, 3 (2018), 1234–1242.

[9] GAO, Y., ZHANG, Y., CAO, Z., GUO, X., AND ZHANG, J. Decoding brain states from fmri signals by using unsupervised domain adaptation. IEEE Journal of Biomedical and Health Informatics 24, 6 (2019), 1677–1685.

[10] BERMÚDEZ-CHACÓN, R., MÁRQUEZ-NEILA, P., SALZMANN, M., AND FUA, P. A domain-adaptive two-stream u-net for electron microscopy image segmentation. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (2018), IEEE, pp. 400–404.

[11] BATESON, M., DOLZ, J., KERVADEC, H., LOMBAERT, H., AND AYED, I. B. Source-free domain adaptation for image segmentation. arXiv preprint arXiv:2108.03152 (2021).

[12] VS, V., VALANARASU, J. M. J., AND PATEL, V. M. Target and task specific source-free domain adaptive image segmentation. arXiv preprint arXiv:2203.15792 (2022).

[13] WANG, J., ZHANG, L., WANG, Q., CHEN, L., SHI, J., CHEN, X., LI, Z., AND SHEN, D. Multi-class asd classification based on functional connectivity and functional correlation tensor via multi-source domain adaptation and multi-view sparse representation. IEEE transactions on medical imaging 39, 10 (2020), 3137–3147.

[14] [14] CHENG, B., LIU, M., SHEN, D., LI, Z., AND ZHANG, D. Multi-domain transfer learning for early diagnosis of alzheimer's disease. Neuroinformatics 15, 2 (2017), 115–132.