CS577: Deep Learning Project

Group 21

The MetaTeacher Framework for Multi-Label Medical Image Classification

Z.Wang,M.Ye,X.Zhu,L.Peng,L.Tian,Y.Zhu, Metateacher: Coordinating multi-model domain adaptation for medical image classification, in: Advances in Neural Information Processing Systems

(https://openreview.net/forum?id=AQd4ugzALQ1)

Github Repository: https://tinyurl.com/4pde5b52

Sukhmani Sandhu (A20513431)
Arpita Jadhav (A20523353)
Saurabh Rajput (A20523129)

Contents

- Motivation
- Introduction
- Datasets
- Problem Statement
- Methods
- Algorithm
- Results and Discussions
- Work that remains
- Conclusions
- References

Motivation

- Medical image analysis is a class of research that requires progressive strides in terms of data acquisition and then classification.
- The problem is not only lack of datasets– but the dearth of annotated data.
- There is also the need for a method to analyse even though the data presented is not in one standard format. This causes a domain shift from training data to unseen (test) data.
- Our project aims to assess the implementation of the Metateacher framework (reference paper)– where we mainly focus on meta-learning and adaptive domain adaptation.
    - This work is a ***first attempt*** at **Multi-Source-Free** and **Semi-Supervised Domain Adaptation** in transfer learning.

Introduction

Domain Shift Challenge & Domain Adaptation

- This refers to the **difference in data distributions** between the source and target domains. It can manifest in various ways, such as changes in lighting conditions, imaging devices, demographics, scan acquisition protocols used by a particular source (laboratory)
    - In healthcare, since access is highly controlled → we have to model a source-free solution
- **Domain-Invariant Representations:** The goal is to learn representations of the data that are insensitive to domain-specific variations. By doing so, the model can generalize well across different domains.
- Limited multi-source domain adaptation(**MSDA**) solutions- mostly all require access to the source domain data

Introduction

Domain Shift & Domain Adaptation

- There exists extensive study on Source-Free Domain Adaptation(**SFDA**) but they use single source which creates:
    - Domain-bias: Hospitals/labs deal with different populations→different class label/frequencies– so existing methods cannot assess the credibility of source domain model with different labels (not quite transferable)
    - If we transfer each source source domain model to the target domain and average predictions? Cannot reveal the complementary information between sources.
- Employ knowledge distillation- multiple teachers, single student scheme. **Meta-learning** involves training a model on multiple tasks, making it more adaptive to new tasks. In domain adaptation, meta-learning can help a model quickly adapt to a new domain with limited labeled data.

Introduction

Combining the best parts of current work

- Shallow and Deep UDA approaches based on feature transformation, domain instance weighting– *need access to source domain data*
- Source-free DA– generative and pseudo-labelling approaches but *only single source domain case*.
- Multi-source DA– approached with distribution alignment and adversarial learning; but current work is shallow models and *requires access to source domain data*.
- So, for better domain alignment we have semi-supervised DA which assumes small number of labeled samples in the target domain.

Introduction

“Multi-Teacher, Single Student” Transfer Scenarios

3 core components

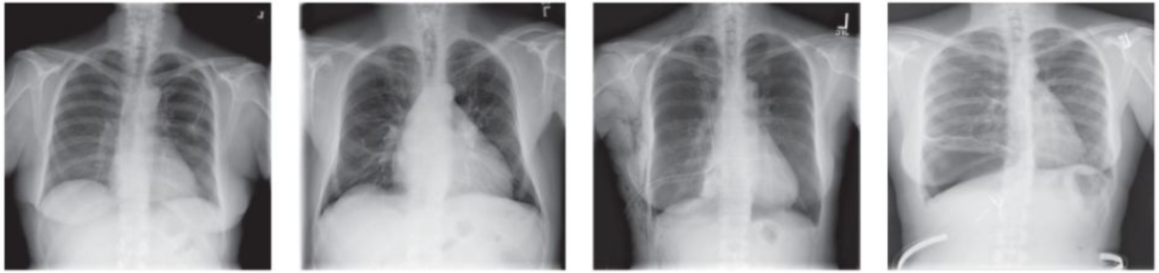
- Models vary in reliability; so we consider **bilevel optimization** strategy to update weights for teachers and student. This is critical for solving the low supervision challenge because it helps leverage collaboration of source models.
- How it’s done: Each teacher model is pretrained. Student model is initialized by a RANDOM teacher.
- We use **coordinated weight learning** to provide different directions for update in multiple teachers (depending on their weight contributions)
- We also have **mutual feedback** from the student

TRANSFER SCENARIOS: NIH-CXR14, CheXpert, MIMIC-CXR to Open-i; NIHCXR14, CheXpert, MIMIC-CXR to Google-Health CXR; CheXpert, MIMIC-CXR to NIH-CXR14; NIHCXR14, CheXpert to Open-i. *This is formatted as from teacher1,teacher2. . . teacherN to student*

Looking at the data

- Chest x-rays are a frequent and cost-effective imaging modality used to diagnose a large number of health conditions because it covers a large area.
- Paper uses five large CXR datasets- open for research access. (We have the NIH CXR14 and Open-i sets)

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



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

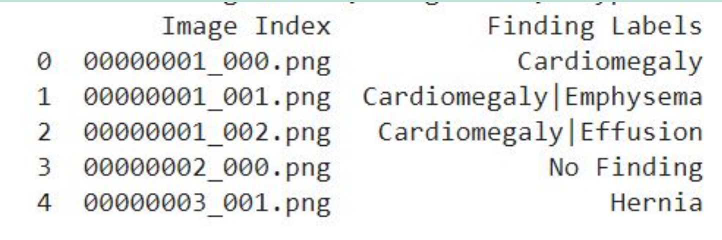
Looking at the data

- Datasets: **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)
- We are currently at a reduced implementation to test our knowledge from the course. We procured the NIH-CXR14 dataset(approximately 112,000 de-identified chest x-rays in PNG)

Looking at the data

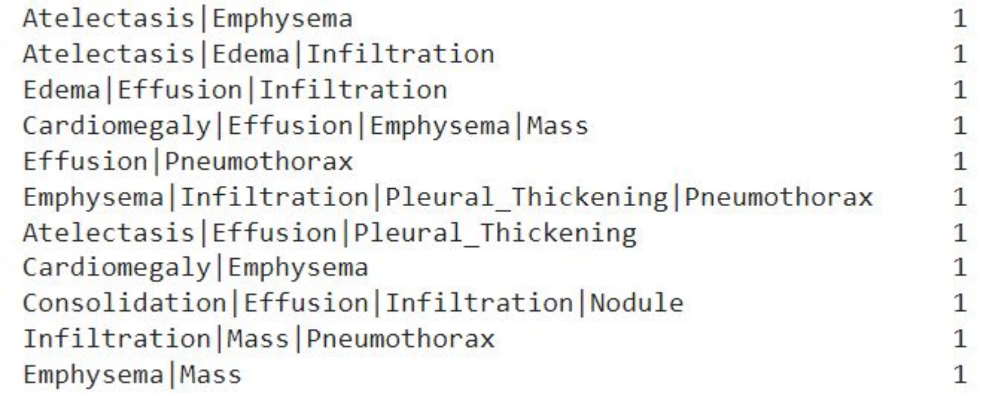
NIH-CXR14...

how the images are read/stored

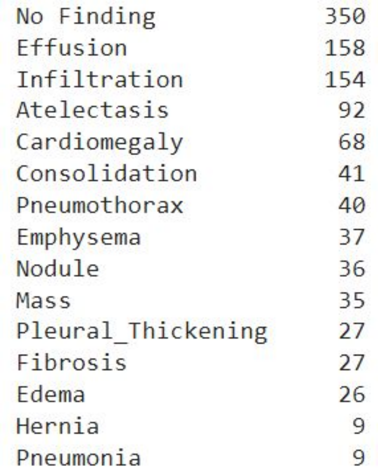


```
        Image Index          Finding Labels
0  00000001_000.png            Cardiomegaly
1  00000001_001.png  Cardiomegaly|Emphysema
2  00000001_002.png   Cardiomegaly|Effusion
3  00000002_000.png              No Finding
4  00000003_001.png                  Hernia
```

the counts for co-occurring in initial small sample



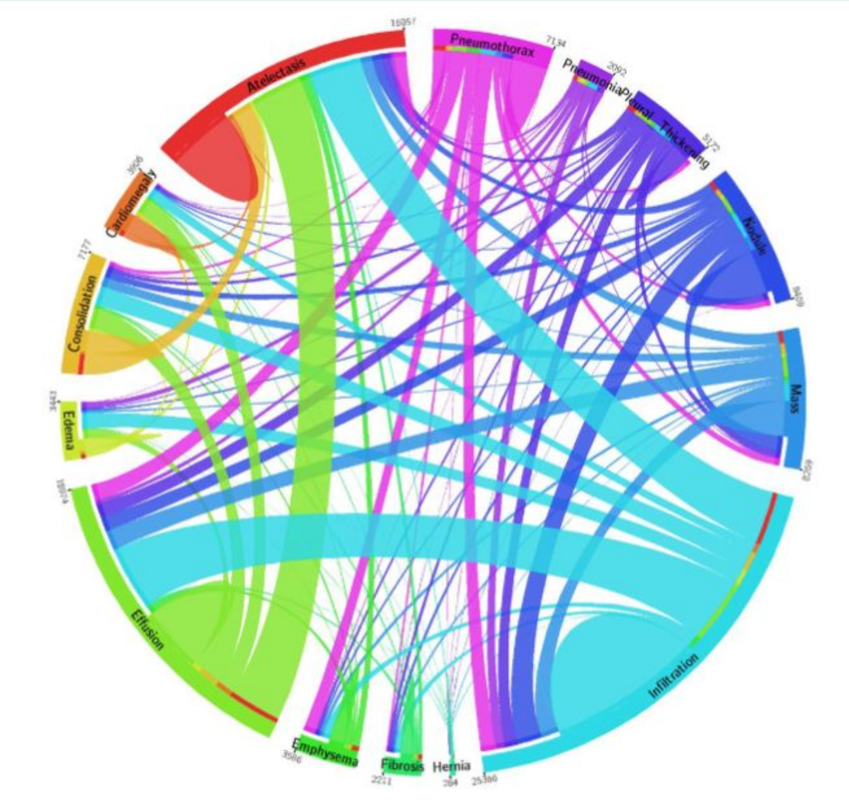
```
Atelectasis|Emphysema                                        1
Atelectasis|Edema|Infiltration                               1
Edema|Effusion|Infiltration                                  1
Cardiomegaly|Effusion|Emphysema|Mass                         1
Effusion|Pneumothorax                                        1
Emphysema|Infiltration|Pleural_Thickening|Pneumothorax       1
Atelectasis|Effusion|Pleural_Thickening                      1
Cardiomegaly|Emphysema                                       1
Consolidation|Effusion|Infiltration|Nodule                   1
Infiltration|Mass|Pneumothorax                               1
Emphysema|Mass                                               1
```



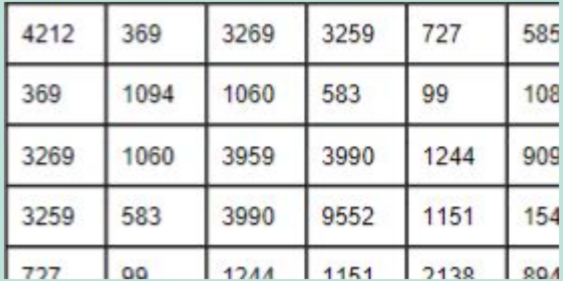
```
No Finding            350
Effusion              158
Infiltration          154
Atelectasis            92
Cardiomegaly           68
Consolidation          41
Pneumothorax           40
Emphysema              37
Nodule                 36
Mass                   35
Pleural_Thickening     27
Fibrosis               27
Edema                  26
Hernia                  9
Pneumonia               9
```

singular counts

Looking at the data



Distributions of the 14 disease categories with co-occurrence statistics



4212	369	3269	3259	727	585
369	1094	1060	583	99	108
3269	1060	3959	3990	1244	909
3259	583	3990	9552	1151	154
727	99	1244	1151	2138	894

Small sample of the co-occurrence matrix

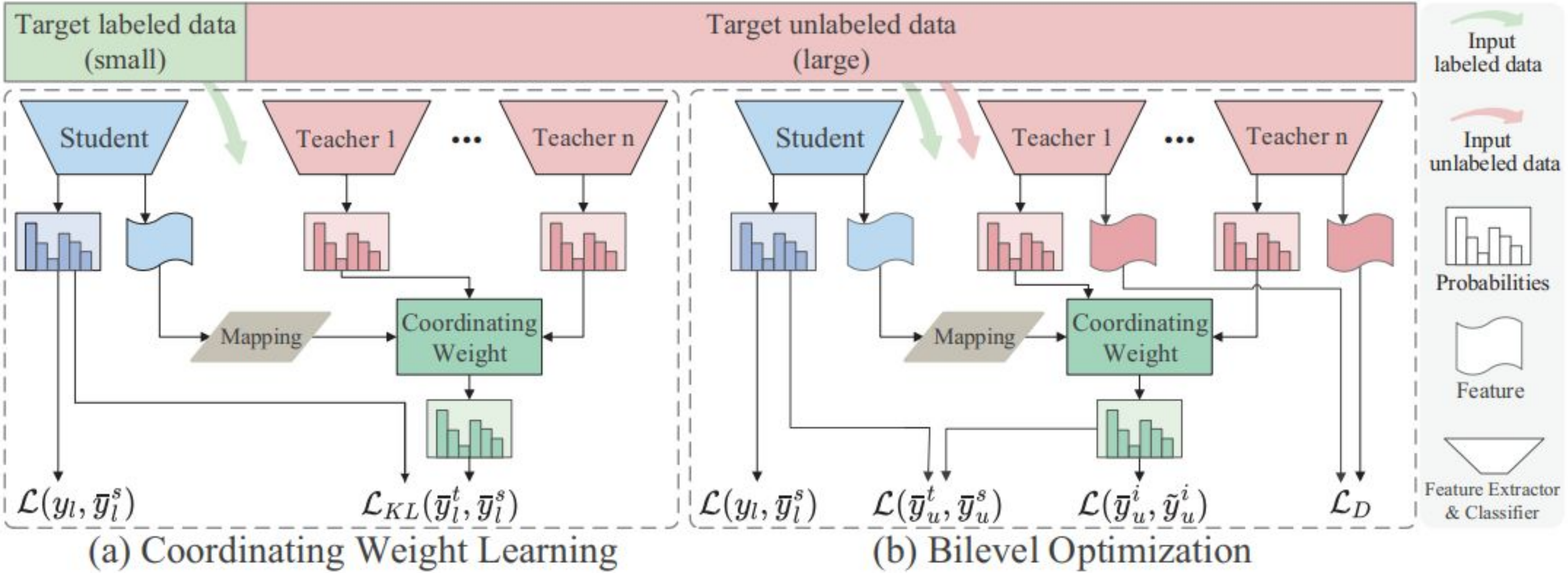
Problem Statement

D_T is the vector(dimension m) for target= $\{(X^t_L, Y^t_L), X^t_U\}$ where Y is the set of label annotations for a small part of the target domain samples X^t_L. Here, X^t_U is the target domain samples that do not have annotations.

D_S(i source domain samples) = $\{(X^i_L, Y^i_L)\}$ where Y is the label for this i-th example X.

For the SMDA to be implemented, as proposed in the paper- when pretrained source classifiers f^T are applied to the target domain, the source dataset 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 mapping such that $f^S: X^t_U \rightarrow Y^t_U$ (predicted classes for target domain points X)


Target labeled data (small)
Target unlabeled data (large)
Input labeled data
Input unlabeled data
Probabilities
Feature
Feature Extractor & Classifier
Student
Teacher 1
Teacher n
Mapping
Coordinating Weight
$\mathcal{L}(y_l, \bar{y}_l^s)$
$\mathcal{L}_{KL}(\bar{y}_l^t, \bar{y}_l^s)$
$\mathcal{L}(\bar{y}_u^t, \bar{y}_u^s)$
$\mathcal{L}(\bar{y}_u^i, \tilde{y}_u^i)$
$\mathcal{L}_D$

(a) Coordinating Weight Learning

(b) Bilevel Optimization

(a) Learning the coordinating weight mapping which will be used to provide guidance for updating the teacher models (b) Alternately updating the teacher and student models. Each teacher is updated with feedback signals from the student and other teachers.

Methods

I) Coordinating Weight Learning

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

- First Part: w.k.t. we pretrain the teacher models on different source domain data- so they present different characteristics → for target domain sample- classification probability of of each teacher model is inconsistent.
- Idea is to optimize each teacher on the target domain samples.
- Which implies the need for the contribution weight of each model to final result.
- As in Fig: To obtain weights- we input the labeled target sample into the student network, get ft. map **B** from feature extraction network **g**. We use pooling to retain significant information. We normalize to get weight matrix **W**.
- Use some mathematical manipulation and then train W.
- Initialize student network with this loss:

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

Methods

$$\Gamma_u(\theta_{T_1}, \cdots, \theta_{T_n}, \theta_S) = \mathcal{L}(\bar{y}_u^t, \bar{y}_u^s)$$

II) Bi-level Optimization

$$\Gamma_l\left(\theta_{T_1}, \cdots, \theta_{T_n}, \theta_S\right) = \mathcal{L}\left(y_l, \bar{y}_l^s\right)$$

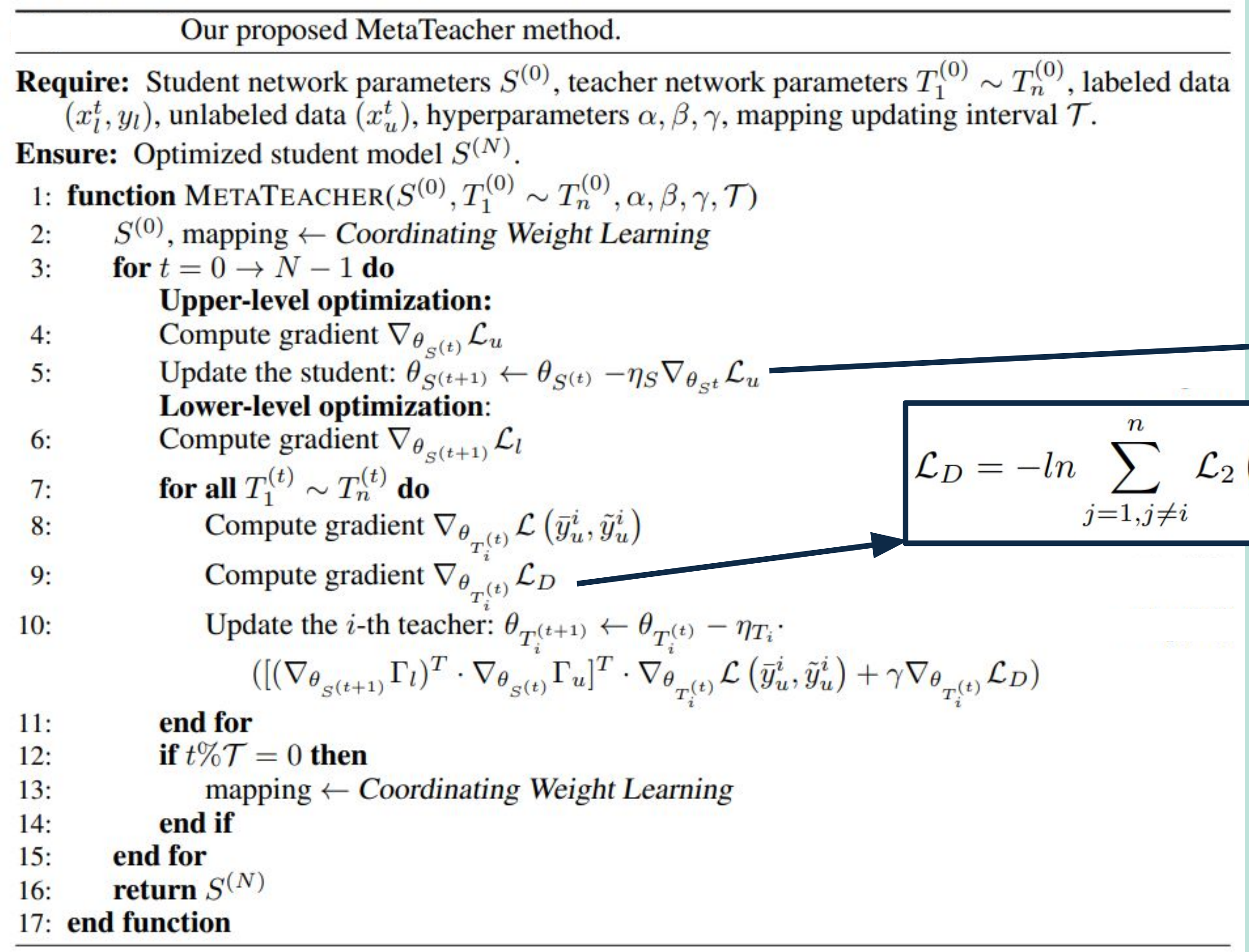
- Second Part: Consists of **(A)** Lower Level Optimization and **(B)** Upper Level Optimization task. **A is a constraint on B.** (Updating the parameters of Teachers and Student is the task objective)
- Here, **B[student]** provides *feedback signals* to **A[teachers]** through performance on labeled data/ weight mapping
- Since tasks are mutually constrained- the teacher parameters cannot be updated until student parameters are optimum → we use **meta-learning** (one step approximation)
- Mathematical manipulations to achieve Divergence Loss and other update rules for respective teacher network.

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

The Algorithm

In order to build a proof sketch, we first put down the assumptions/conditions that the paper works with:

a) Multiple source domain models are trained on multi-label image datasets
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



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 $S^{(N)}$.

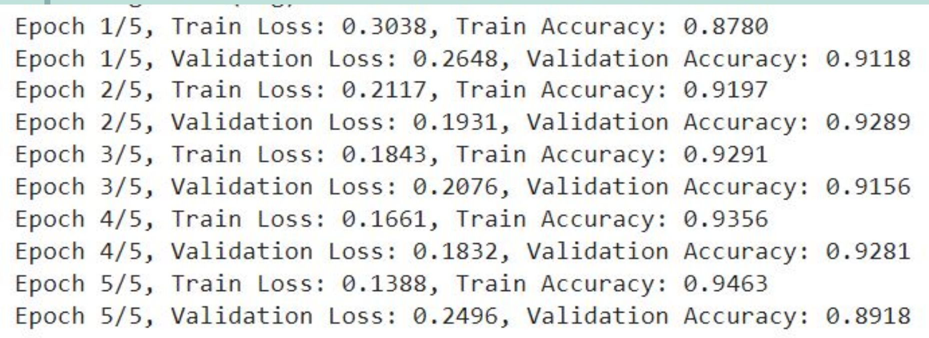
```
function MetaTeacher($S^{(0)}, T_1^{(0)} \sim T_n^{(0)}, \alpha, \beta, \gamma, \mathcal{T}$)
    $S^{(0)}$, mapping $\leftarrow$ Coordinating Weight Learning
    for $t = 0 \rightarrow N-1$ do
        Upper-level optimization:
        Compute gradient $\nabla_{\theta_{S^{(t)}}} \mathcal{L}_u$
 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

$$\theta'_S = \theta_S - \eta_S \cdot \nabla_{\theta_s} \Gamma_u,$$

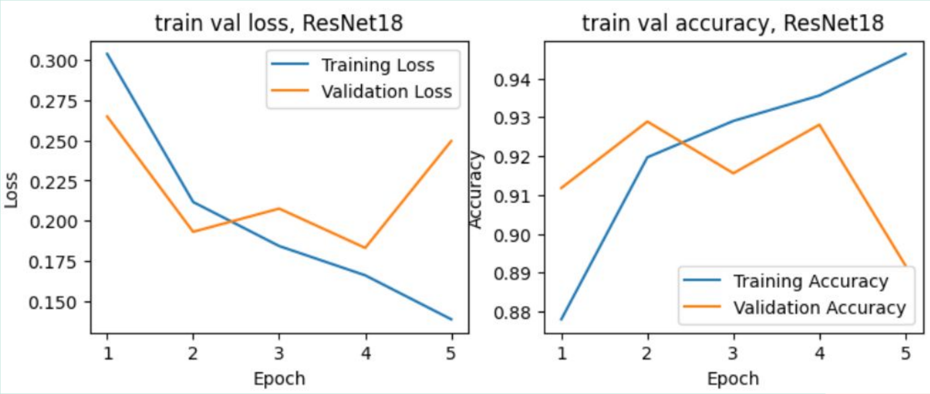
$$\mathcal{L}_D = -ln \sum_{j=1, j \neq i}^{n} \mathcal{L}_2\left(B_{T_i}\left(x_u^t; \theta_{T_i}\right), B_{T_j}\left(x_u^t; \theta_{T_j}\right)\right)$$

Results and Discussions

- At this stage, we are working to implement the algorithm on our own simplified transfer scenario: **NIHCXR14(a), NIHCXR14(b) to Open-i**
  Here, we break the large NIH dataset into two. ***TEACHERS→ AlexNet, DenseNet***
  STUDENT→ ResNet18
- Intermediate stage: working with models, “breaking the NIH dataset into very small sizes to train the student/teacher models”
- Example results on ResNet18 model:

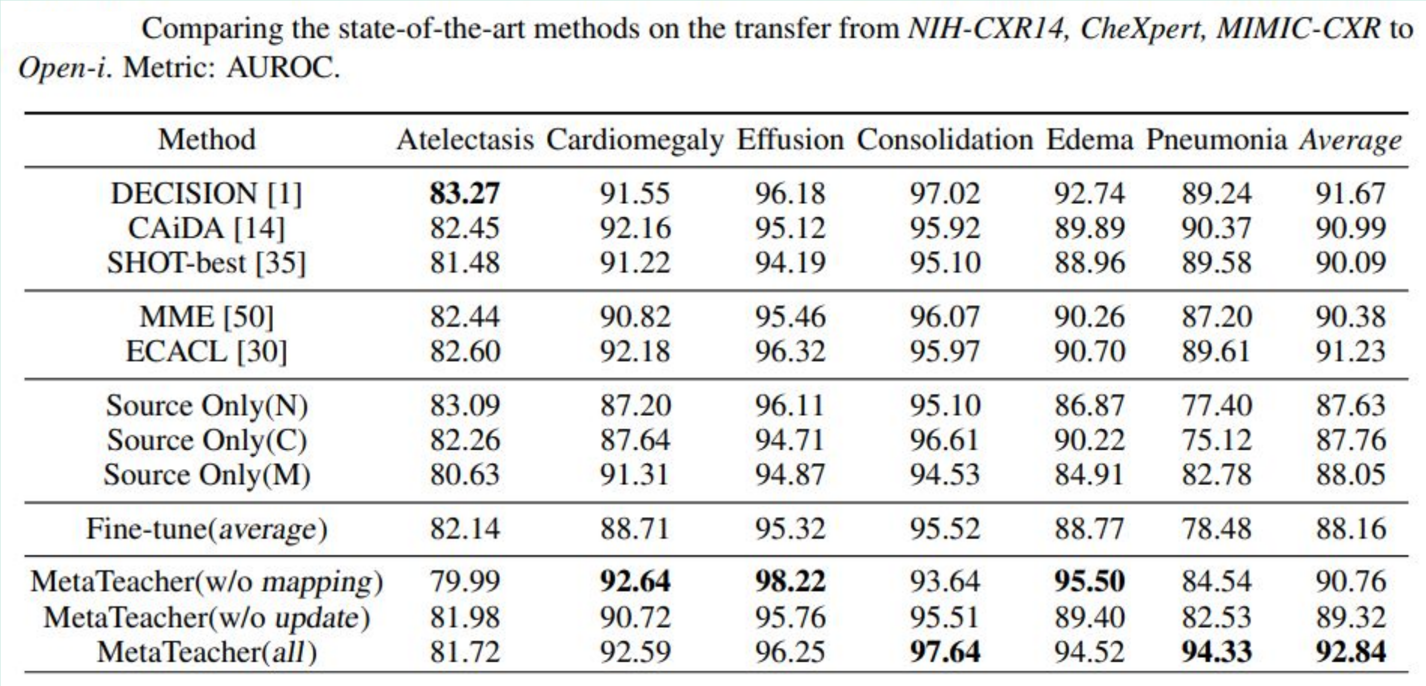


```
Epoch 1/5, Train Loss: 0.3038, Train Accuracy: 0.8780
Epoch 1/5, Validation Loss: 0.2648, Validation Accuracy: 0.9118
Epoch 2/5, Train Loss: 0.2117, Train Accuracy: 0.9197
Epoch 2/5, Validation Loss: 0.1931, Validation Accuracy: 0.9289
Epoch 3/5, Train Loss: 0.1843, Train Accuracy: 0.9291
Epoch 3/5, 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
```



Results and Discussions

(The reference paper):



Comparing the state-of-the-art methods on the transfer from *NIH-CXR14*, *CheXpert*, *MIMIC-CXR* to *Open-i*. Metric: AUROC.

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**

Conclusions

- We notice that performance is still competitive without coordinated weight learning: why?
  - for student update: averaging over teacher predictions is beneficial
  - fixed W is involved in teacher optimization
  - ie. BiLevel Optimization is indispensable in terms of overall gain, than Coordinated Weight Learning.
  - But Coordinated Weight Learning will judge which disease label the teacher is best at, by weight. So we end up selecting results from predictions of the best teachers, for each category.
- Parameters: α gives optimal performance at 0.5, any lower and W is not trained well → optimization direction for teacher can't be determined. If β too large → can't express the relationship between source domains(CWL ineffective). If 0, overfitting. γ influences divergence loss.

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 Xray Database and Benchmarks on Weakly- Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462- 3471,2017

THANK YOU :)