Semi-supervised Classification of Diagnostic Radiographs with NoTeacher: A Teacher that is Not Mean

Balagopal Unnikrishnan[1], Cuong Manh Nguyen[1], Shafa Balaram[1,2], Chuan Sheng Foo[1], and Pavitra Krishnaswamy[1(✉)]

[1] Institute for Infocomm Research, A*STAR, Singapore
{balagopalu,pavitrak}@i2r.a-star.edu.sg
[2] National University of Singapore, Singapore

Abstract. Deep learning approaches offer strong performance for radiology image classification, but are bottlenecked by the need for large labeled training datasets. Semi-supervised learning (SSL) methods that can leverage small labeled datasets alongside larger unlabeled datasets offer potential for reducing labeling cost. However, few studies have demonstrated gains of SSL for real-world radiology image classification. Here, we adapt three leading SSL methods (Mean Teacher, Virtual Adversarial Training, Pseudo-labeling) for radiograph classification, and characterize their performance on two public X-Ray and CT classification benchmarks. We observe that Mean Teacher can achieve good performance gains in the low labeled data regime, but is sensitive to hyperparameters and susceptible to confirmation bias. To address these issues, we introduce a novel SSL method named NoTeacher. This method incorporates a probabilistic graphical model to maximize mutual agreement between student networks, thereby eliminating the need for a teacher network. We show that NoTeacher outperforms contemporary SSL baselines by enforcing better consistency regularization, and achieves over 90% of the fully supervised AUROC with less than 5% labeling budget.

Keywords: Semi-supervised deep learning · Classification · Multi-label · X-rays · CT · Mean teacher

1 Introduction

Deep learning approaches offer state-of-the-art performance for a range of image classification applications in radiology. Recent successes include chest radiograph diagnosis, brain tumor prognostication, fracture detection, and breast cancer screening [7,11,14,21]. However, these efforts typically require large labeled

B. Unnikrishnan and C. M. Nguyen—Equal Contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-59710-8_61) contains supplementary material, which is available to authorized users.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12261, pp. 624–634, 2020.
https://doi.org/10.1007/978-3-030-59710-8_61

training datasets assembled through resource-intensive labeling by specialized domain experts. Further, as discordance between expert raters can result in noisy labels, even more resource-intensive consensus rating across multiple blinded raters is often required [6].

One way to reduce this labeling burden is to use semi-supervised learning (SSL) approaches. These approaches leverage large numbers of unlabeled images alongside smaller numbers of labeled images for model development. The best performing approaches are typically *consistency-based*; these encourage a classifier's predictions to be consistent with a target on the unlabeled data. For instance, the popular Mean Teacher (MT) method [26] enforces consistency between predictions from two networks, termed the student and teacher, where the teacher is a time-averaged version of the student and is used for inference at test time. MT and other SSL approaches have been applied to computer vision classification benchmarks and to radiology image segmentation [3,10,29]. By contrast, few studies have demonstrated gains of these SSL methods for radiology image classification.

Here, we focus on semi-supervised learning for the under-studied radiograph (X-Ray and CT) classification task that typically involves detection of multiple abnormality types in the same image. We adapt three leading semi-supervised deep learning methods (Pseudo-labeling [9], Virtual Adversarial Training [15], and Mean Teacher [26]) to this multi-label setting and characterize performance using a realistic semi-supervised annotation and evaluation process. We observe that MT offers good performance gains in the low labeled data regime, but note its sensitivity to hyperparameters and vulnerability for confirmation bias. These MT limitations are in part due to reliance on a time-averaged student as a consistency target (teacher) that essentially leads the model to enforce self-consistency. To address these issues, we introduce the NoTeacher method that instead enforces consistency between two independent networks. Our method is derived by marginalizing out a latent consistency variable in a probabilistic graphical model. Results on the NIH-14 Chest X-Ray and RSNA Brain Hemorrhage CT datasets [4,22] show that NoTeacher outperforms competing methods on both datasets with minimal hyperparameter tuning, and achieves over 90% of the fully supervised AUROC with less than 5% labeling budget.

Related Work: We briefly introduce the methods characterized in this work. Pseudo-labeling (PSU) [9] is a classic SSL algorithm that is simple and commonly used. It is an iterative algorithm where one trains a model on labeled data, uses this trained model to infer pseudo-labels on the unlabeled data, then includes these pseudo-labels to enlarge the training set in the next iteration. Virtual Adversarial Training (VAT) [15] regularizes the output posterior distribution to be isotropically smooth around each input image. Mean Teacher (MT) [26] enforces consistency between two networks: a student model and a teacher model whose parameters are the exponential moving average (EMA) of the student's.

Aside from these methods that we characterized, other SSL methods have been proposed. For example, MixMatch [2] is a consistency-based SSL method which emphasizes data augmentation. However, the Mix-Up augmentation

technique cannot be directly applied to medical images. Further, GAN-based methods [23] learn the underlying distribution from unlabeled data, but do not easily scale to high resolution images typical in radiology [12,13] and patch-wise adaptations [8] neglect the necessary global context. Deep Co-training [20] combines consistency and adversarial training in a multi-view framework, but the official implementation is not released and the method is not easily adapted to our multi-label setting. Finally, most related is GraphXNet [1], a graph-based label propagation method for X-Ray classification, but it does not support the multi-label setting and was not compared to other SSL methods.

2 Methods

Adaptations for Multi-label Classification: Radiology images may be associated with more than one label, where each label presents a binary classification task. Therefore, in our experiments, we used multi-task neural networks, trained using the sum of binary cross-entropy losses, one per label. The adaptation of VAT for our multi-label setting is described in the Supplement.

Mean Teacher Background: To provide context for our NoTeacher method, we briefly review Mean Teacher [26]. We consider a semi-supervised single-label classification task with image input x and binary output $y \in \{0, 1\}$. Mean Teacher (MT) employs two networks with identical architecture: a student model F_S and a teacher model F_T. A schematic illustration of the MT model is provided in Fig. 1 (a). Given a batch $\mathbf{x}$ of training data, MT applies random augmentations η_S, η_T to generate inputs $\mathbf{x}_S$ and $\mathbf{x}_T$ for the corresponding model. During the feed-forward pass, MT computes a total loss combining the usual supervised cross-entropy (CE) classification loss and a consistency loss (mean-squared error MSE on the posteriors):

$$\mathfrak{L}_{\text{MT}} = \text{CE}\left(\mathbf{y}, \mathbf{f}_S^{\text{L}}\right) + \lambda_{\text{cons}} \text{MSE}\left(\mathbf{f}_S, \mathbf{f}_T\right), \tag{1}$$

where $\mathbf{f}_S, \mathbf{f}_T$ are posterior outputs from the student and teacher networks, $\mathbf{f}_S^{\text{L}}$ is the student's posterior output on the labeled data, and λ_{cons} is a consistency weight hyperparameter. The student model is updated directly by backpropagation using gradients of the loss $\mathfrak{L}_{\text{MT}}$. Meanwhile, the teacher model updates its parameters by computing an EMA over the student's parameters. MT improves over supervised learning when the teacher generates better expected targets, or pseudo labels, to train its student. Recent papers have adapted MT for medical imaging tasks such as MR segmentation [19,29] and nuclei classification [24].

However, because the teacher model is essentially a temporal ensemble of the student model in the parameter space, MT has two potential drawbacks. First, enforcing consistency of the student model with its historical self may lead to confirmation bias or unwanted propagation of label noise [5]. Second, the teacher model is sensitive to the choice of the EMA hyperparameter, causing performance degradation when this hyperparameter is set outside a narrow range, as seen in realistic evaluation regimes [18]. Moreover, since MT does not establish a systematic method to compute the consistency weight λ_{cons}, the process to

tune this hyperparameter for varied datasets is unclear. Several variants of MT therefore try to enforce consistency using other regularization schemes [19,24].

NoTeacher Overview: To address the above challenges, we introduce a new method. Figure 1(b) illustrates the overall framework of NoTeacher (NoT), where we have made two major changes: (a) we removed the EMA update so that the networks are completely detached and (b) we trained the model with a novel loss function $\mathfrak{L}_{\text{NoT}}$ based on a probabilistic graphical model to enhance consistency. Since the two networks are treated equally, we index them numerically as F_1 and F_2.

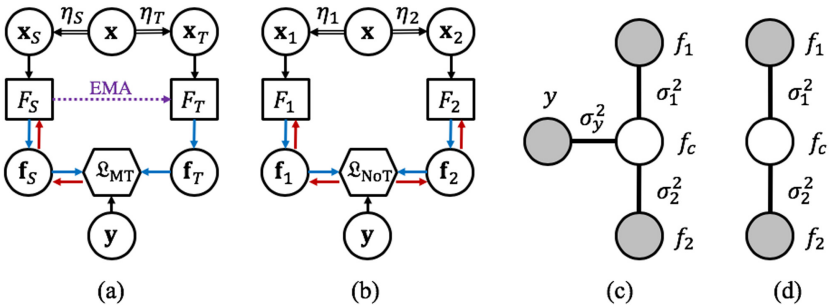

Fig. 1. Training procedure of (a) MT and (b) NoT for one iteration of semi-supervised learning. Double-line arrows denote random data augmentations, purple dotted arrow represents EMA update, blue and red arrows are forward and backward passes, respectively. NoT graphical model on a single (c) labeled and (d) unlabeled image.

NoTeacher Graphical Model and Loss: For each image x, the networks F_1, F_2 take as inputs two different images x_1, x_2 generated by applying random augmentations η_1, η_2, respectively. Since x_1, x_2 are generated from the same image, the network outputs $f_1 = F_1(x_1)$ and $f_2 = F_2(x_2)$ should be similar. Furthermore, if the image is labeled, then those outputs should also match the label. Because label y is binary, the network outputs can be interpreted as *posteriors* of the label, e.g., $f_1 = \Pr(y = 1|x_1)$. Inspired by previous works in semi-supervised regression [16] and kernel learning [30], we consider y, f_1, f_2 as random variables and design an undirected graphical model to impose probabilistic constraints on them. Figure 1 (c) and (d) show the NoT graphical models for labeled and unlabeled images respectively. The observed variables y, f_1, f_2 are represented by separate nodes, each is connected only to a latent variable called the *consensus function* $f_c \in [0, 1]$. As its name implies, f_c enforces the mutual agreement of the posteriors on both labeled and unlabeled data. When the label is available, f_c acts as an information relay between the posteriors and y. For analytical tractability, the differences $f_1 - f_c$ and $f_2 - f_c$ are assumed to follow Gaussian distributions $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$ respectively. To account for labeling noise, we also assume the difference $y - f_c$ follows a Gaussian distribution $\mathcal{N}(0, \sigma_y^2)$.

Given a training batch of n_{L} labeled and n_{U} unlabeled images, the likelihood can be expressed as follows

$$\begin{aligned} p\left(\mathbf{y}|\mathbf{x}\right) \propto &\exp\left(-\lambda^{\mathrm{L}}_{y,1}\|\mathbf{f}^{\mathrm{L}}_1 - \mathbf{y}\|^2 - \lambda^{\mathrm{L}}_{y,2}\|\mathbf{f}^{\mathrm{L}}_2 - \mathbf{y}\|^2\right) \cdot \\ &\exp\left(-\lambda^{\mathrm{L}}_{1,2}\|\mathbf{f}^{\mathrm{L}}_1 - \mathbf{f}^{\mathrm{L}}_2\|^2 - \lambda^{\mathrm{U}}_{1,2}\|\mathbf{f}^{\mathrm{U}}_1 - \mathbf{f}^{\mathrm{U}}_2\|^2\right), \end{aligned} \tag{2}$$

where $\mathbf{f}^{\mathrm{L}}_{\bullet}, \mathbf{f}^{\mathrm{U}}_{\bullet}$ are vectors containing the posteriors on labeled and unlabeled data, respectively, $\mathbf{y}$ is the vector of labels, and $\lambda^{\mathrm{L}}_{y,1}, \lambda^{\mathrm{L}}_{y,2}, \lambda^{\mathrm{L}}_{1,2}, \lambda^{\mathrm{U}}_{1,2}$ are derived as detailed in Supplement. Maximizing the likelihood in (2) yields the following loss function:

$$\begin{aligned} \mathfrak{L}_{sq} = &\lambda^{\mathrm{L}}_{y,1}\|\mathbf{f}^{\mathrm{L}}_1 - \mathbf{y}\|^2 + \lambda^{\mathrm{L}}_{y,2}\|\mathbf{f}^{\mathrm{L}}_2 - \mathbf{y}\|^2 \\ &+ \lambda^{\mathrm{L}}_{1,2}\|\mathbf{f}^{\mathrm{L}}_1 - \mathbf{f}^{\mathrm{L}}_2\|^2 + \lambda^{\mathrm{U}}_{1,2}\|\mathbf{f}^{\mathrm{U}}_1 - \mathbf{f}^{\mathrm{U}}_2\|^2. \end{aligned} \tag{3}$$

To avoid vanishing gradients when using sigmoid activations with a squared loss, on the labeled data, we apply CE loss on the posteriors instead. Our final NoT loss is therefore

$$\begin{aligned} \mathfrak{L}_{\mathrm{NoT}} = &\lambda^{\mathrm{L}}_{y,1}\mathrm{CE}\left(\mathbf{y}, \mathbf{f}^{\mathrm{L}}_1\right) + \lambda^{\mathrm{L}}_{y,2}\mathrm{CE}\left(\mathbf{y}, \mathbf{f}^{\mathrm{L}}_2\right) \\ &+ \lambda^{\mathrm{L}}_{1,2}\mathrm{MSE}\left(\mathbf{f}^{\mathrm{L}}_1, \mathbf{f}^{\mathrm{L}}_2\right) + \lambda^{\mathrm{U}}_{1,2}\frac{n_{\mathrm{U}}}{n_{\mathrm{L}}}\mathrm{MSE}\left(\mathbf{f}^{\mathrm{U}}_1, \mathbf{f}^{\mathrm{U}}_2\right), \end{aligned} \tag{4}$$

where we set $n_{\mathrm{L}} = n_{\mathrm{U}}$ to cancel out the fraction in the last term. These changes also enable a fair comparison with MT in our experiments. The first two terms of $\mathfrak{L}_{\mathrm{NoT}}$ represent the supervised losses while the last two terms enforce mutual agreement between the classifiers, thus enhancing consistency of the predictions.

3 Experiment Setup

Datasets: We now describe the two datasets used in our experiments, and provide statistical breakdowns in the Supplement.

NIH-14 Chest X-Ray [28]*:* The first dataset we used comprises 112,120 frontal chest X-Ray images, where images are annotated for presence of one or more of 14 pathologies (14 binary labels). 53.9% of images are normal (negative for all 14 labels) and 46.1% abnormal; 40.1% of abnormal images are positive for more than one pathology. We used publicly available training (70%), validation (10%) and testing (20%) splits with no patient overlaps [31].

RSNA Brain CT [4]*:* The second dataset we used is a collection of 19,530 CT brain exams from the Stage 1 training dataset of the RSNA 2019 Challenge, where images are annotated for presence of one or more of 5 types of intracranial hemorrhage (5 binary labels). We focus on slice-level classification, and consider only one study per patient. 85.8% of images are normal (negative for all 5 labels) while the remaining 14.2% are positive for bleeds; 30.1% of the

abnormal images are positive for multiple bleeds. We derived random training (60%), validation (20%) and testing (20%) splits with no patient overlaps. We obtained pre-processed images from [17]. The pre-processing steps used are in accordance with leading solutions in the RSNA Challenge [25] and include: (a) converting the raw pixel values to Hounsfield Units using the slope and rescale intercept from the original DICOM files, (b) windowing to restrict all pixel values to the width around the center as reported in the individual DICOM files, (c) normalizing to range [0, 255], and (d) resizing to 256×256.

Design of Realistic Semi-supervised Experiments: We model a realistic annotation process where clinical raters extract a finite pool of N images for model development, and proceed to label them systematically. First, we consider a labeling budget of $L = L_T + L_V$ images, where L_T denotes the labeled subset of the training data, and L_V denotes the labeled validation set. A practical labeling budget implies that $L_T \gg L_V$ with L_V very small. Second, for a given L, we need to randomly sample a subset for labeling from the overall pool of N images. Implicitly, this means the subset of L images cannot be chosen based on their labels i.e., it is infeasible to fit any stratified class distribution requirements. For lower budgets, this constraint requires repeatedly sampling and labeling images until representative numbers are obtained for each class. Third, increases in L require maintaining the existing labeled images, and progressively adding on images and labeling them (implicitly decreasing unlabeled set size).

For our experiments, we had to align the above process to publicly available data collections that were already split into training, validation and test sets. Hence, for a given labeling budget L, we randomly sampled L images proportionately from the training and validation splits. Then, for the unlabeled set, we only considered the remaining portion of the training split to ensure that the unlabeled validation split does not inform training. All our experiments assume a fixed-size held out test set and do not count the test set definition as part of labeling budget for model development. The above practical requirements preclude the conveniences of large validation sets, stratified labeling, balanced class distributions, fixed unlabeled budgets that are often encountered in standard SSL benchmarking papers [18], and introduce additional challenges.

Implementation Details: We describe the supervised training setup and hyperparameter tuning procedures. SSL validation details are in Supplement.

Supervised Training Backbone: We use the same backbone network architecture for supervised (SUP) and semi-supervised methods to ensure fair comparisons: DenseNet121 for NIH-14 Chest X-Ray as per [22] and DenseNet169 for RSNA Brain CT (used by the leading RSNA Challenge solution [25]). In each case, we initialized the network with weights pretrained on ImageNet. For train-time augmentations, we employed random horizontal flipping, resizing, and center cropping. We normalized all images based on ImageNet statistics, and used the Adam optimizer (learning rate 10^{-4}, $\beta = [0.9, 0.999]$, $\epsilon = 10^{-8}$, and weight decay 10^{-5}).

Hyperparameter Tuning: We tuned hyperparameters in accordance with literature norms for all semi-supervised methods to ensure fair comparisons. In particular, we tuned ϵ for VAT, EMA decay and consistency weight for MT, and considered variations required for different labeling budgets. For NoT, even though there are four weights, only the ratio between the weights is of consequence – hence only the ratio hyperparameter requires tuning. As the two networks have similar architecture, we selected $\sigma_1^2 = \sigma_2^2$ and varied the labeling noise σ_y^2. The tuning process and final parameter choices are provided in Supplement.

Parameter Averaging: In order to enable fair comparison with MT, we keep an EMA copy for the supervised baseline and VAT; and report the best result from either the trained model or the EMA copy. This way, performance gains reported are not just because of averaging but also due to the consistency mechanism.

4 Results

We systematically evaluate performance of each SSL algorithm as a function of varying labeling budget L. We start L at 500 images of the development dataset or the number required to have at least 1 positive image per label (whichever is higher). For each labeling budget, we also evaluate performance of a comparable supervised baseline (SUP) trained purely on the labeled images. All experiments maintain the same held-out test set for rigorous comparison. As both classification objectives are multi-label tasks, we compute the per-class AUROC and report average across all classes.

Performance vs. Labeling Budget: Figures 2 and 3 show results on the NIH-14 X-Ray and RSNA Brain CT datasets respectively. Detailed performance numbers are provided in Supplement.

For low labeling budgets, (a) the semi-supervised methods offer strong gains over the supervised baseline, and (b) our NoT method outperforms the other semi-supervised methods. To surpass 90% of the fully supervised AUROC, NoT requires less than 5% labeling budget for NIH-14 X-Ray dataset and less than 2.5% labeling budget for RSNA CT dataset. First, For NIH-14, with 5% labeling budget, NoT gains 6.4% over the corresponding supervised baseline and over 3.1% vs. other SSL methods. NoT also outperforms the comparable GraphXNet result at 5% labeling budget [1]. For RSNA CT, with 2.5% labeling budget, NoT gains 3.7% over the supervised baseline (SUP) and over 2.6% vs. other SSL methods. Second, for higher labeling budgets, performance of all semi-supervised methods converges, suggesting saturation of gain from unlabeled data. Across all methods, NoT can achieve over 90% of the fully supervised AUROC with less than 5% labeling budget. Third, at 5% labeling budget for NIH-14 and 2.5% labeling budget for RSNA CT, we compute the AUROC gain of each SSL method over the corresponding supervised baseline (SUP). We plot these gains for each class in a heatmap format, where conditions are ordered by number of images in the labeled set. For rarer conditions with lower prevalence in the labeled subsets, NoT gains much more than other SSL methods.

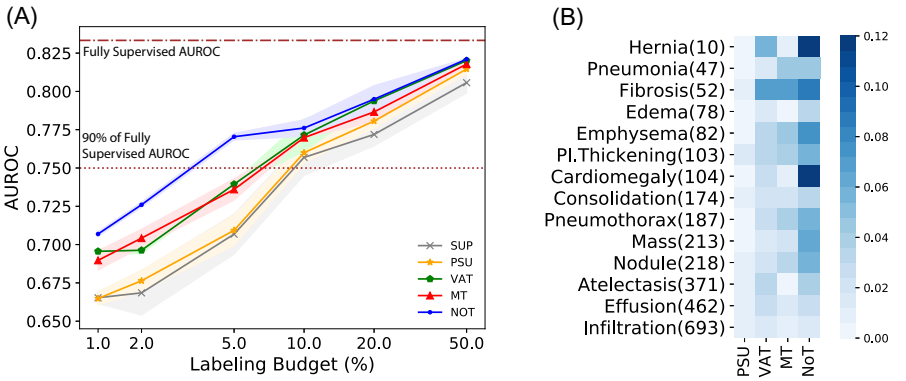

Fig. 2. Performance Results for the NIH-14 Chest X-Ray Dataset.(A) Average AUROC vs. Labeling Budget. AUROC evaluated with 1177, 1569, 3923, 7846, 15693, 39234 labeled images (with L_T : L_V set to 70:10). Fully supervised baseline (dash-dotted line) is based on 78468 images. (B) Class-wise SSL vs. SUP AUROC gains for 5% labeling budget. Hernia (10) indicates 10 images with hernia in labeled set.

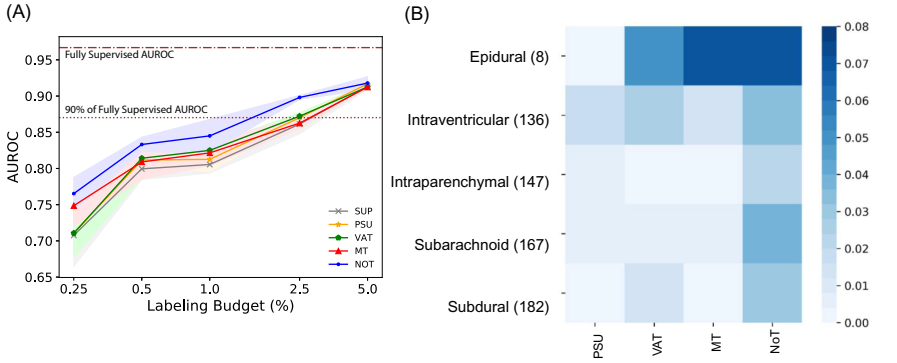

Fig. 3. Performance Results for the RSNA Brain CT Dataset. (A) Average AUROC vs. Labeling Budget. AUROC evaluated with 749, 1777, 3495, 6744, and 17242 images. Labeling budget set at scan level and selected scans have all slices labeled. L_T : L_V set to 60:20. Fully supervised baseline (dash-dotted line) is based on 352839 images. (B) Class-wise SSL vs. SUP AUROC gains for 2.5% labeling budget. Epidural (8) indicates 8 images with epidural bleed in labeled set.

Connections to Co-training and Label Propagation: We posit that these gains arise from the multi-view formulation of NoT. While VAT, PSU and even MT are essentially single-network models (the MT teacher network is learned passively via EMA), NoT is a multi-view learning technique which benefits from having multiple views of the data. Being a co-training method, NoT also has connections with label propagation [27].

Performance Analyses of NoT vs. MT: To understand how NoT improves performance over MT, we train both models on the NIH-14 X-Ray dataset with a 5% labeling budget and save the predictions on validation data. We compare consistency between the student-teacher networks of MT and the consensus between networks F_1 and F_2 of NoT by reporting the *disagreement* count or the number of validation images with different predictions (Fig. S1 in Supplement). On average, NoT reduces disagreement count by 51.87 % compared to MT. In addition, after the first 400 iterations, NoT maintains an AUROC variance of 8.39×10^{-5}, while MT shows a much higher variance of 3.42×10^{-4}.

5 Conclusion

We adapted and characterized three leading semi-supervised methods for multi-label radiograph classification using a realistic annotation and evaluation process. To further improve the best of these methods, MT, we introduce the NoTeacher method (NoT) to better enforce consistency and reduce confirmation bias. We demonstrate that NoT provides strong performance gains on two public X-Ray and CT classification benchmarks, and achieves over 90% of the fully supervised AUROC with less than 5% labeling budget. Our results suggest feasibility for deep learning with minimal supervision on radiology images and provide a strong benchmark for future developments.

Acknowledgements. Research efforts were supported by funding and infrastructure for deep learning and medical imaging research from the Institute for Infocomm Research, Science and Engineering Research Council, A*STAR, Singapore. We acknowledge insightful discussions with Jayashree Kalpathy-Cramer and Praveer Singh at the Massachusetts General Hospital, Boston USA. We also thank Ashraf Kassim from the National University of Singapore for his support.

References

1. Aviles-Rivero, A.I., Papadakis, N., et al.: GraphX$^{\mathrm{NET}}$ - chest X-ray classification under extreme minimal supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 504–512 (2019)
2. Berthelot, D., et al.: Mixmatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 5050–5060 (2019)
3. Feng, Z., et al.: Semi-supervised learning for pelvic MR image segmentation based on multi-task residual fully convolutional networks. In: IEEE International Symposium on Biomedical Imaging, pp. 885–888 (2018)
4. Flanders, A.E., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol. Artif. Intell. **2**(3), e190211 (2020)
5. Ke, Z., et al.: Dual student: breaking the limits of the teacher in semi-supervised learning. In: IEEE International Conference on Computer Vision, pp. 6728–6736 (2019)

6. Langlotz, C.P., et al.: A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology **291**(3), 781–791 (2019)
7. Lao, J., et al.: A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Sci. Rep. **7**(1), 1–8 (2017)
8. Lecouat, B., et al.: Semi-supervised deep learning for abnormality classification in retinal images. In: NeurIPS Machine Learning for Health Workshop (2018)
9. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning, vol. 3 (2013)
10. Li, X., et al.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. In: IEEE Transactions on Neural Networks and Learning Systems (2020)
11. Lindsey, R., et al.: Deep neural network improves fracture detection by clinicians. Proc. Natl. Acad. Sci. **115**(45), 11591–11596 (2018)
12. Madani, A., et al.: Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In: IEEE International Symposium on Biomedical Imaging, pp. 1038–1042 (2018)
13. Madani, A., et al.: Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. Nat. Partner J. Digit. Med. **1**(1), 1–11 (2018)
14. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
15. Miyato, T., et al.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2018)
16. Nguyen, C.M., et al.: Partial Bayesian co-training for virtual metrology. IEEE Trans. Ind. Inform. **16**(5), 2937–2945 (2019)
17. Oh, R.: RSNA Train/Test png (256 × 256). Kaggle dataset. https://kaggle.com/richul/rsna_png_128_128
18. Oliver, A., Odena, A., Raffel, C.A., et al.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems, pp. 3235–3246 (2018)
19. Perone, C.S., Cohen-Adad, J.: Deep semi-supervised segmentation with weight-averaged consistency targets. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS - 2018. LNCS, vol. 11045, pp. 12–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_2
20. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 142–159. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_9
21. Rajpurkar, P., Irvin, J., et al.: Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. Pub. Libr. Sci. Med. **15**(11), e1002686 (2018)
22. Rajpurkar, P., Irvin, J., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
23. Salimans, T., et al.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)

24. Su, H., Shi, X., Cai, J., Yang, L.: Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 559–567. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_62
25. Tao, S.: RSNA intracranial hemorrhage detection. GitHub repository (2019). https://github.com/SeuTao/RSNA2019_Intracranial-Hemorrhage-Detection
26. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, pp. 1195–1204 (2017)
27. Wang, W. and Zhou, Z.H.: A new analysis of co-training. In: International Conference on Machine Learning, pp. 1135–1142 (2010)
28. Wang, X., et al.: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3462–3471 (2017)
29. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
30. Yu, S., et al.: Bayesian co-training. J. Mach. Learn. Res. **12**, 2649–2680 (2011)
31. Zech, J.: Reproduce-chexnet. GitHub repository (2018). https://github.com/jrzech/reproduce-chexnet