Source-free Few-shot Semi-supervised Domain Adaptation for CT Image Classification

Bowen Song
ICME
Stanford University
bowen18@stanford.edu

Abstract

The application of deep learning techniques in medical imaging is emerging. However, the medical data is hard to obtain and may be subject to privacy and confidential issues. The labeling process can also be costly due to the requirement of domain knowledge. It is crucial to develop semi-supervised algorithms that can be trained with very few data. In this work, we explore algorithms that can perform a basic but important task: classify the body part of a CT image. We consider the most challenging scenario: only few training images are available and most of them are unlabeled (one-shot paired data with 9-shot unlabelled data). We propose a psuedo-label based source-free semi-supervised domain adaptation algorithm to tackle this problem. Our result shows significant improvement over the baselines.

1. Introduction

Deep learning models have achieved state-of-the-art performance on medical imaging applications such as tumor classification and organ segmentation [10]. However, the training heavily relies on a large amount of training data (e.g., thousands of examples), which is very difficult to obtain in medical imaging due to privacy concerns and confidentiality [1]. The labeling process is also very expensive since it requires well-trained medical professionals to determine the correct label. On the other hand, different scanner has different settings which gives very different CT images with various noise level. In additional to the few data problem, we also need to deal with the domain shift problem [1]. In this work, we propose to address the challenge above by utilizing a semi-supervised pseudo-label based algorithm. We consider one of the most challenging scenarios: source-free adaptation, which means that we can not access any data from the source domain that we are using. This has a lot of practical meaning since in medical applications, we usually want to quickly train a model without accessing any external data. There exists a lot of source-free domain adaptation works that attempts to minimize the entropy at test time [16] [15] . We also introduce a few-shot semi-supervised setting so that we only have a one-shot three-way labeled images with another 27 unlabeled images. This simulates a realistic clinical setting that we have few experienced doctors that can label the image correctly. There are lots of existing works on few-shot learning [12][4][9][14] [6]. Among these methods, meta learning and metric learning are some of the most popular ones. However, these methods require accessing a lot of images from other datasets for pretraining. In our work, we want to adapt an existing model to a novel test-time environment with very limited data available. Hence, the problem we are solving is a very challenging one with lots of limitation. We propose a pseudo-label propagation algorithm that effectively address this problem. The idea is based on that the pretrained model itself encodes some knowledge in the predictions, and we can use that knowledge to create pseudo-labels as a data augmentation technique, and then increase the accuracy further. Our work shows significant improvements over the fine-tuning baseline. One advantage of our algorithm is that we do not need any data from the source domain, which is a requirement for many meta-learning based algorithms[9].

2. Related Work

2.1. Source-free Domain Adaptation

In many scenarios, deep learning models suffer from performance degradation, when generalized to a new testing domain out of training data distribution. Unsupervised domain adaptation (UDA) aims for improving model performance in the target domain without the ground truth labels, where the testing target domain and training source domain are different with a distribution shift [16]. Unsupervised domain adaptation methods often align the feature distribution between the two domains in order to learn domain-invariant features [7].

Test-time unsupervised adaptation further restricts the access to the training data from the source domain, and adapt the model to test-time data based on the assumption of the distribution of the softmax probabilities [13] or changing the parameters in the pretrained model [2]. In practice, the test-time data distribution may be continually changing, which makes traditional UDA methods not suitable since the assumption that test-time data comes from a consistent distribution is violated. [13] proposed a teacher-student model with psuedo-label augmentation and stochastic model restoration. Most of these methods are based on the assumption of separability of softmax probabilities or invariant feature for domain shift.

2.2. Few-shot Learning

Given abundant training examples for the base classes, few-shot learning algorithms aim to learn to recognize novel classes with a limited amount of labeled examples [14]. The easiest way is to perform fine-tuning, which uses the pretrained model as initialization and retrain the model. There are three most popular methods to perform few-shots learning: initialization based, metric learning based, and hallucination based methods. Initialization based methods tackle the few-shot learning problem by "learning to fine-tune", which learns the hyperparameter of the neural network [4] [12]. One approach aims to learn good model initialization (i.e., the parameters of a network) so that the classifiers for novel classes can be learned with a limited number of labeled examples and a small number of gradient update steps. Another line of work focuses on learning an optimizer. While these initialization based methods are capable of achieving rapid adaptation with a limited number of training examples for novel classes, [2] shows that these methods have difficulty in handling domain shifts between base and novel classes. Distance metric learning based methods learns a metric such that similar images are close to each other and unsimilar images are far away from each other [6][**?**][8]. Some contrastive losses are applied to address this issue. Hallucination based methods learns data augmentation to compensate the limited training data size. This class of methods learns a generator from data in the base classes and use the learned generator to hallucinate new novel class data for data augmentation. However, to train the generators also needs accessing the base classes, and applying the style change at test-time, which may not be available in clinical practices. In summary, the problem we propose to solve is more challenging than traditional few-shots learning since we are not allowed to access the source-domain data.

2.3. Untrained Network Prior

There have been a lot of studies on neural network structure contains implicit prior knowledge. [11] shows that untrained CNN fits the clean image first and noise later. With the help of early stopping, untrained CNN can achieve competitive performance on tasks such as denoising and super-resolution. [5] further applies untrained CNN for compressed sensing tasks with learned regularization. However, there are few studies about using the untrained network prior on classification tasks, especially in the few-shot setting.

3. Methods

3.1. Problem statement

The goal of our work is to adapt a pretrained model to the few-shot learning problem of identifying the body part a CT scan image. Specifically, let g_θ be the pretrained model, x_0, y_0 be the one-shot labeled target domain data, and x_1 be the unlabeled target domain data. Our goal is to find a model that gives the highest accuracy of classification. We assume that we can access the parameters of the pretrained model.

he objective of our training is

$$\min_{\phi} \sum_{i}^{D} l(f_\phi(g_\theta, x_i), y_i)$$

x is the CT image, y is the corresponding body part and D is the size of the training set. We want to utilize the prior information from the pretrained model and the training set to learn from the few-shot training data.

3.2. Untrained network prior for few-shot classification

In order to leverage the network prior of CNN, we propose to introduce an early stopping when fine tuning the pretrained model. Specifically, the fine tuning objective is given by

$$\arg\min_{\phi} \sum_{i}^{D} CE(f_\phi(g_\theta, x_i), y_i)$$

, where CE is the cross-entropy loss. Consider a sliding window of the training loss of a few epochs, at the starting of the training, the training loss tends to decrease rapidly, and then flattens and converges. However, when it is close to converge, the loss may oscillates a lot. Hence, we propose a metrics d. such that

$$d = \frac{VAR(l_{train})}{E[l_{train}^2]}$$

, where l_{train} is the training loss of a sliding window of epochs. We propose to use $\arg\min_{epoch} d$ as the stopping point of the training. This will avoid the model to become overfitted to the few-shot training set, and learn an effective representation.

3.3. Psuedo-label propogation

We propose to further leverage the prior knowledge in the pretrained model for improving the performance of classification. We assume that the pretrained model itself encodes some knowledge about the images, and this knowledge is also shared after fine-tuning. Specifically, when the prediction gives a high confidence (probability), that means the model is very certain about this prediction, hence, we can create a psuedo-label pair that consists of the image and the prediction. Then we can retrain the model again using the newly added image psuedo-label pair. Our algorithm is listed below.

Algorithm 1 Source-free Few-shot Semi-supervised CT Classification

Input: labeled data x_0, y_0, unlabeled data x_1, pretrained model g_θ, confidence level t, number of iterations *iter*, learning rate λ

while There are new psuedo-labels **do**
  for i from 0 to *iter* **do**
    $\phi \leftarrow \phi - \lambda \nabla_\phi \sum_j^D CE(f_\phi(g_\theta, x_{0j}), y_{0j})$
  end for
  $\hat{d} \leftarrow \arg\min_{train} \frac{VAR(l_{train})}{E[l_{train}^2]}$
  $\phi \leftarrow \phi_{\hat{d}}$
  for $x \in x_1$ **do**
    if $f_\phi(x) > t$ **then**
      Append x and $f_\phi(x)$ to the training set
    end if
  end for
end while

3.4. Alternative methods

We also try to train only the top layer of the pretrained model while freezing weights of all other layers. We also introduce an entropy loss for unsupervised learning $l_{CE} = \sum_j \sum_{i \in C} log(f_\phi(x_j)_i)$. The idea is the a successful model should be able to have high confidence in prediction, hence the entropy should be minimized. This approach is widely used in many UDA works [16]. However, these methods do not overperform the fine-tuning baseline in our work.

4. Data Collection

We collected the full-dose CT data from the LDCT (Low dose CT images) [3]. The data consists of full-dose projection, full-dose reconstructions, low-dose projection and low-dose reconstruction. We select full-dose CT reconstructions from 10 patients for our study. For each patient, we collect 3 body parts (head, chest and abdominal). For each patient, the data from one body part consists of multiple 2D slices that can be combined into a 3D image. For the purpose of this study, we randomly sample 10 slices from each body part for each patient, and normalize the data into the range of [0,1] (the original range is from -1000 to 1400. For training, we use two slices from the first patient for each of the three body parts (two-shot three-way), all remaining 2D slices are left for training. We also augmented the dataset by simulating the noisy measurement scenario (FBP reconstruction when the measurement contains noise).

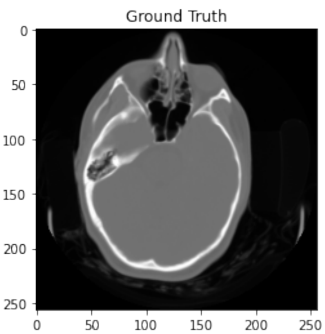

Figure 1. An example of our collected data

5. Experiments

5.1. Classification Accuracy

To study the performance of our proposed method, we use the data from the first patient as the training set and the data from the other nine patients as the testing set. We use the first slice of each organ of the patient as the labeled data, and the rest of the slices as unlabeled data. The baseline model without pretraining and the fine-tuning baseline is trained on the labeled data (3 labeled images). Our proposed algorithm is run on both the labeled and unlabeled data (3 labeled images and 27 unlabeled images). We use three pretrained model for conducting our experiments: VGG-16, ResNet-50, and ResNet18. The performance of different algorithms is reported in terms of accuracy on the test set (270 images) as in the table below. We set the number of iterations be 100, the learning rate for first step finetuning as 0.01; for second step and beyond of psuedo-label propogation be 0.001, and the confidence level to be 0.8.

Table 1. Results

Method	VGG-16	ResNet-50	ResNet-18
No-pretrain	0.51	0.59	0.57
Fine-tuning	0.77	0.82	0.81
Psuedo-label Propogation (Ours)	**0.85**	**0.93**	**0.90**

5.2. Ablation Study

In this section, we study the effect of early stopping and psuedo-label propogation. Even though it is a well-known phenomena of untrained CNN prior in the image denoising studies, untrained network prior is a relative unknown subject for classfication. We found that the most significant boost to the performance is due to the psuedo-label propogation. Early stopping with the criteria d makes our model more robust to overfitting.

Table 2. Results

Method	VGG-16	ResNet-50	ResNet-18
FT	0.77	0.82	0.81
FT + ES	0.78	0.82	0.80
FT + PS	0.84	**0.93**	**0.91**
FT + PS + ES	**0.85**	**0.93**	0.90

5.3. Analysis

We found that our proposed psuedo-label propogation method significantly improves the few-shots classification accuracy. The early stopping technique, on the contrary does not seem that significant. However, in case of overfitting happens the early stopping method can effectively reduce the overfitting by detecting the onset of overfitting. From figure 2, we can clearly visualize the trend of d during training (reaches the minimum in the middle of training due to oscillation of overfitting).

Figure 2. the d metrics of training curve for resnet-50 and resnet-18

6. Conclusion

We presented a simple yet effective way for Source-free Few-shot Semi-supervised Domain Adaptation for CT Image Classification. By utilizing the knowledge from the pretrained model such as confidence level and training curve shape, we are able to improve the performance of fine-tuned model on the test set with limited data. We also make the model more robust to overfitting by introducing a detection metric. Our results show significant improvement over the baseline fine-tuning methods for all 3 pretrained models. However, our results are still limited since we do not included the most state-of-the-art few-shot semi-supervised methods for comparison due to difficulty of implementation, which will be left for future work. In sumamry, we method is effective for classify the body type from one CT image with few training data with most data being unlabeled. This should offer more help to physicians in the clinical practice for streamlining processing of CT images.

References

[1] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. 1

[2] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 2

[3] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013. 3

[4] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 1, 2

[5] C. Kim, J. Lee, and J. Shin. Zero-shot blind image denoising via implicit neural representations. *arXiv preprint arXiv:2204.02405*, 2022. 2

[6] G. Koch, R. Zemel, R. Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015. 1, 2

[7] J. Liang, D. Hu, and J. Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2

[8] I. Melekhov, J. Kannala, and E. Rahtu. Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 378–383. IEEE, 2016. 2

[9] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 1

[10] L. Shen, W. Zhao, and L. Xing. Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. *Nature biomedical engineering*, 3(11):880–888, 2019. 1

[11] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 2

[12] J. Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018. 1, 2

[13] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 2

[14] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. 1, 2

[15] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, and S. Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 1

[16] H. Zhang, Y. Zhang, K. Jia, and L. Zhang. Unsupervised domain adaptation of black-box source models. *arXiv preprint arXiv:2101.02839*, 2021. 1, 3