Knowledge Distillation
for Semi-supervised Domain Adaptation

Mauricio Orbes-Arteainst[1,2,3,4(✉)], Jorge Cardoso[4], Lauge Sørensen[1,3],
Christian Igel[1], Sebastien Ourselin[4], Marc Modat[4], Mads Nielsen[1,3],
and Akshay Pai[2]

[1] Department of Computer Sceince, University of Copenhagen,
Copenhagen, Denmark
[2] Cerebriu A/S, Copenhagen, Denmark
[3] Biomediq A/S, Copenhagen, Denmark
[4] King's College London, London, UK
henry.m.orbes_arteaga@kcl.ac.uk

Abstract. In the absence of sufficient data variation (e.g., scanner and protocol variability) in annotated data, deep neural networks (DNNs) tend to overfit during training. As a result, their performance is significantly lower on data from unseen sources compared to the performance on data from the same source as the training data. Semi-supervised domain adaptation methods can alleviate this problem by tuning networks to new target domains without the need for annotated data from these domains. Adversarial domain adaptation (ADA) methods are a popular choice that aim to train networks in such a way that the features generated are domain agnostic. However, these methods require careful dataset-specific selection of hyperparameters such as the complexity of the discriminator in order to achieve a reasonable performance. We propose to use knowledge distillation (KD) – an efficient way of transferring knowledge between different DNNs – for semi-supervised domain adaption of DNNs. It does not require dataset-specific hyperparameter tuning, making it generally applicable. The proposed method is compared to ADA for segmentation of white matter hyperintensities (WMH) in magnetic resonance imaging (MRI) scans generated by scanners that are not a part of the training set. Compared with both the baseline DNN (trained on source domain only and without any adaption to target domain) and with using ADA for semi-supervised domain adaptation, the proposed method achieves significantly higher WMH dice scores.

Keywords: Semi-supervised learning · Domain adaptation ·
Knowledge distillation · White matter hyperintensities

1   Introduction

In the presence of a large training dataset that covers all possible data variations, deep neural networks (DNNs) can achieve super-human performance in image recognition and semantic segmentation tasks. However, in medical image segmentation tasks large annotated training datasets are often scarce. In addition, training and test data are drawn from different distributions. For example, the images were obtained using different scanners at different sites or the demographics of the subjects differ. This violation of the i.i.d. assumption (i.e., that training and test data are drawn independently from the same distribution) typically has the effect that the performance on the test data is significantly worse than on the training data.

Domain adaptation (DA) approaches try to alleviate the problem of applying models in new domains with different characteristics. In particular, semi-supervised DA methods provide a way to learn structure from unlabeled data in new domains. Among the several semi-supervised DA (SSL-DA) methods proposed, the most popular one is *adversarial training based domain adaptation* (ADA). ADA relies on generating features that are invariant with respect to a domain discriminator. ADA requires extensive parameter optimization due to the necessity of a robust discriminator. And a recent study pointed out the flaws in the evaluation of SSL-DA methods [1].

In this paper, we evaluate a modified *knowledge distillation* (KD) [2,3] method for generalizing DNNs to new domains with a common clinical problem in contrast to using ADA methods. The datasets chosen for evaluation not only involve different magnetic resonance images (MRIs), but also were acquired on subjects with different demographic makeup. Through our evaluation, we show that the proposed KD is generally able to achieve better dice scores in segmenting white matter hyperintensities (WMH) on datasets that are not a part of the training data and do not share any attributes when compared to baseline and ADA.

2   Related Work

Among the recent works on DA, several methods rely on using a small amount of data (*annotated*) to fine-tune a baseline model [4,5]. The performance of this approach not only relies on a new – albeit small – set of annotations but also on the choice of the set. In contrast, SSL-DA do not use data annotations on new target domains. Adversarial training is a popular SSL-DA method [6–8]. Here, networks are trained in such a way that the generated features are agnostic to the data domain with respect to a domain discriminator. A similar solution, ADA, was employed by [9] to adapt networks to be agnostic to domain changes.

Another class of DA method use KD to transfer representations between data domains. For instance, [10] proposed using KD to transfer knowledge between different modalities of the same scene. Closely related to our work is [11], where the authors propose to use omni-supervised learning (OSL) to include unlabelled data in the learning process. Here, data distillation is used to generate an

ensemble of predictions from multiple transformations of unlabeled data, using a teacher model, to generate new training annotations. The proposed method differs from this method on two accounts: (a) Only soft labels are used to train the single student network, where the idea is to improve segmentation by learning label similarities from unannotated data (b) the data included in the training of the student involves data from new domains in small amounts in contrast to OSL.

3   Methods

In SSL-DA methods, we assume the source domain images and their annotations, $(x_s, y_s) \in \mathbf{X}_s$, are drawn from a distribution $p_s(x_s, y)$. The target domain images $x_t \in \mathbf{X}_t$, are drawn from a distribution $p_t(x_t, y)$ where there are no annotations available. We consider classification into K classes. In an ideal scenario, where p_s and p_t are sufficiently similar, the goal is to find a feature representation mapping f that maps an input to K scores, where the ith score models (up to a constant) the logarithm of the probability that the input belongs to class K. These scores can then be mapped by $\sigma : \mathbb{R}^K \to \mathbb{R}^K$ to probability maps over the classes. SSL-DA first finds a function f_s performing well on a source domain and then finds a new f_t based on f_s that performs well on the target domain. Vanilla supervised learning methods rely on including annotations from both $\mathbf{X}_s$ and $\mathbf{X}_t$.

In the popular ADA method, the goal is to minimize the distance between the empirical distributions of $p_s(f_s(\mathbf{X}_s)|y)$ and $p_t(f_t(\mathbf{X}_t)|y)$. Here, a discriminator D is a neural network that distinguishes between the two domains. Therefore, the discriminator acts as a discrepancy measure that brings the two distributions together. Overall, adversarial training involves train a network that generates f in a standard supervised manner that is indistinguishable by a discriminator [6,9].

3.1   Knowledge Distillation for Domain Adaptation

KD [2] was originally intended to compress neural networks with high number of parameters with networks of lower complexity. The objective is to teach a simpler student network to imitate a more complex trained teacher network, through a loss function called the distillation loss. To perform unsupervised domain adaptation, we proposed to use the teacher/student learning strategy. Specifically, the data from the source domain is used to train a teacher model in a supervised fashion. Then, the trained teacher is used to generate posterior probability maps or soft labels on the union of source and target data. These posterior probabilities are used instead of usual hard labels to train the student or target model. Note, this approach can take advantage of large amounts of unlabeled data acquired from any number of domains. An attractive feature of distillation loss is the soft representation of one-hot encoded label vectors which allow the student to be optimized over a smoother optimization landscape. Moreover, the smooth

representation of labels also allows the learning of label similarities, which is particularly useful in learning boundaries in semantic segmentation tasks. The proposed semi-supervised learning method is formulated below.

Training the Teacher or Source Domain Model: Consider a set of N manually annotate images from a source domain $\mathbf{X}_s = \{(x_i, y_i), i = 1 \ldots N\}$, where $x_i \in \mathbb{R}^d$ represent a d-dimensional MR scan, with $v = 1 \ldots V$ voxels, and $y_i \in [0,1]^K$ with $\|y_i\|_1 = 1$ its correspondent label. Assuming there is a set F_s that holds functions $f : \mathbb{R}^d \to \mathbb{R}^K$ we aim to learn a feature representation f_s (teacher model) which follows the optimization of a loss function, l, according to Eq. (1)

$$\underset{f \in F_s}{\arg\min} \frac{1}{N} \sum_{x_i \in \mathbf{X_s}} l(y_i, \sigma(f_s(x_i))) \tag{1}$$

$$[\sigma(z)]_k = \frac{\mathrm{e}^{[z]_k}}{\sum_{l=1}^{K} \mathrm{e}^{[z]_l}} \tag{2}$$

In a standard supervised learning way, the teacher network is optimized using the cross-entropy loss function (or any differentiable loss function of choice).

Training the Student or Target Model: Even though f_s is suitable to segment the images from the source domain $\mathbf{X}_s$, it may not be suitable for data coming from a different data distribution $\mathbf{X}_t$. Our goal is find a function $f_t \in F_t$, which is suitable to segment data from $\mathbf{X}_t$. Assuming, we have access to a limited set of unlabeled scans in the target domain $\mathbf{X}_t = \{x_i, i = 1 \ldots M\}$, we can then create a set

$$\mathbf{X}_U = \{(x_i, y_i) \,|\, x_i \in \mathbf{X}_s, y_i = f_s(x_i), 1 \leq i \leq N\} \cup$$
$$\{(x_i, y_i) \,|\, x_i \in \mathbf{X}_t, y_i = f_s(x_i), 1 \leq i \leq M\}$$

that may be used to optimize a student using the distillation loss. Through soft-representations of this union dataset, the student is expected to learn a better mapping to the labels than the teacher network. When training the student network, we consider probability distributions over the labels as targets, not single classes. This representation reflects the uncertainty of the prediction by the teacher network. The function f_t is found by (approximately) solving,

$$\underset{f \in F_t}{\arg\min} \frac{1}{(N+M)} \sum_{x_i \in \mathbf{X_U}} l(\sigma(T^{-1} f_s(x_i)), \sigma(f_t(x_i))) \ , \tag{3}$$

Here, $T > 1$ is the temperature parameter which controls the softness of the class probability prediction given by f_s.

4 Experiments and Results

4.1 Databases

The **WMH segmentation challenge** (https://wmh.isi.uu.nl/) dataset is a public database that contains T1-weighted and FLAIR scans for 60 subjects

from three different clinics. The data also consists of manual annotations of
WMH from presumed vascular origin. T1-weighted images have been registered
to FLAIR since annotations were performed in this space. The images were also
corrected for bias field inhomogenities using SPM12. An important feature of
this dataset is that the scanners and demographics have variance as show in the
Table 1.

Table 1. Summary of data characteristics in the WMH challenge database

Clinic	Scanner name	Voxel size(m^3)	Size	# of images
Utrech	3T Philips Achieva	$0.96 \times 0.95 \times 3.00$	$240 \times 240 \times 48$	20
Singapore	3T Siemens TrioTim	$1.00 \times 1.00 \times 3.00$	$252 \times 232 \times 48$	20
Amsterdam	3T GE Signa HDxt	$1.20 \times 0.98 \times 3.00$	$132 \times 256 \times 83$	20

4.2   Experimental Setup

One of the main objectives of the paper is to use semi-supervised learning to
perform domain adaptation. We use the WMH challenge dataset to perform
cross-clinical experiments in segmenting WMH on FLAIR images. We consider
several scenarios to establish the performances of ADA and KD. The scenarios
are described below. Note that, to evaluate the performance of the algorithms,
dice overlap measures are used throughout.

- Lower bound baseline, **L-bound**: Here a baseline DNN model is trained on the
  source dataset to establish a lower bound performance. The DNN is trained
  on the source domain images henceforth referred to as **S**, and tested on 20
  subjects from a target dataset **T**.
- Upper bound baseline, **U-bound**: Here, a baseline DNN model is trained like
  L-Bound, however, the training dataset is a union of images from both **S** and
  a subset of **T** (10 subjects, with annotations). The network is evaluated on
  the remaining 10 subjects in **T**.
- Adversarial domain adaptation, **ADA**: Following [9], we attempt at training
  a DNN model that is invariant to data domains. In this paper, to be consistent
  with KD, we train the domain discriminator based on the final layer of the
  baseline, in contrast to what was proposed in [9]. We use a discriminator
  composed of 4 convolutional layers with 8, 16 32, 64 number of filters, followed
  by 3 fully connected layers with 64, 128 and 2 neurons. For this experiment,
  like U-bound, the training dataset is a union of images from both **S** and a
  subset of **T** (10 subjects, without annotations). The network is evaluated on
  the remaining 10 subjects in **T**.
- Knowledge distillation, **KD**: The experimental setup for KD is the same as
  ADA. A temperature of 2 is used in the softmax for the distillation loss. The
  student network trained is identical to the teacher network whose architecture
  is a standard UNet (like L-bound, U-bound, and ADA) optimized with an

ADAM loss function and a learning rate of 10^{-4} with is gradual decrease after epoch 150. The network is trained for 400 epochs.

– Adaptation on-the-fly: A clinically relevant scenario is adapting to a small set of test images on the fly by keeping the teacher/baseline model constant. To validate this scenario, we apply ADA and KD on the same 10 unannotated **T** that are included in the training, but subject-wise. In other words, separate adaptation is performed on each instance of **T**, instead of including them together.

4.3   Results

Various combinations of mismatched (in terms of clinics) training and testing data were used. For instance, if the training data is from clinic 1 (Utrecth), the testing data is from either clinic 2 (Singapore), or clinic3 (Amsterdam). We did not test on two different clinics even though this scenario is practical. Table 2 illustrates mean dice coefficients (two folds) for each of the scenarios mentioned in Sect. 4.2 except for adaptation on the fly which is illustrated in Table 3. KD outperformed ADA in nearly all scenarios except for domain adaptation from Singapore clinic to Utrecht clinic and vice versa. For domain adaptation from Utrecht clinic to Singapore clinic, ADA was significantly better than KD. In the vice-versa situation, KD achieved a better mean which is statistically not significant. In all other scenarios, KD yielded statistically better dice overlaps compared to ADA. Note that the statistical comparison are made only between

Table 2. Illustrates dice overlaps (with variance). Bold fond indicates statistical significance at 5%, p-values (paired-sample t-test at was used to computed p-values, which were $0.0002 < p < 0.02$). Only ADA and KD methods are considered in the statistical comparison.

Training	Test			
	Method	Utrech	Singapore	Amsterdam
Utrech	L-bound		0.6126 (0.1092)	0.7207 (0.0793)
	ADA		**0.7004 (0.1057)**	0.7144 (0.0968)
	KD		0.6456 (0.0905)	**0.7548 (0.0755)**
	U-bound		0.8031 (0.1148)	0.7704 (0.0787)
Singapore	L-bound	0.6693 (0.2271)		0.7368 (0.0931)
	ADA	0.6859 (0.2036)		0.7337 (0.0912)
	KD	0.6924 (0.2103)		**0.7499 (0.0877)**
	U-bound	0.7063 (0.2016)		0.7699 (0.0851)
Amsterdam	L-bound	0.6471 (0.2086)	0.6811 (0.1172)	
	ADA	0.6800 (0.2128)	0.7202 (0.1154)	
	KD	**0.6909 (0.2135)**	**0.7482 (0.0975)**	
	U-bound	0.7208 (0.1851)	0.7988 (0.0869)	

Table 3. Mean dice overlaps from the adaptation-on-the-fly scenario. Bold fond indicates statistical significance at 5%, p-values (paired-sample t-test at was used to computed p-values, which were $0.0003 < p < 0.04$). Only ADA and KD methods are considered in the statistical comparison.

Training	Test			
	Method	Utrech	Singapore	Amsterdam
Utrech	KD		**0.6285 (0.097**	**0.7465 (0.0855)**
	ADA		0.7075 (0.095)	0.7220 (0.0995)
Singapore	KD	**0.6945 (0.1825)**		0.7425 (0.0805)
	ADA	0.6680 (0.1945)		0.7370 (0.0880)
Amsterdam	KD	**0.6745 (0.2005)**	**0.7395 (0.1165)**	
	ADA	0.6625 (0.1890)	0.7100 (0.1125)	

ADA and KD. In the adaptation-on-the-fly scenario, KD yields significantly better dice overlaps on a majority of the scenarios, the superior performance of ADA remains in the experiment that involves domain adaptation from Utrecht clinic to Singapore clinic. However, in the vice-versa scenario, KD performance better than ADA. To illustrate the differences in segmentations between KD and ADA, we plot the segmentations (scenario, Utrecht clinic to Amsterdam clinic) in Fig. 1. As illustrated, both the methods perform quite well in segmenting lesions with relatively larger volume, however, the main difference is evident in segmenting smaller lesions, specially in the deep white matter regions. It is interesting to note that the adaptation-on-the-fly and the classical scenarios yield nearly the same dice indicating a good generalisability and less dependency on the choice of the small dataset coming from the target domain.
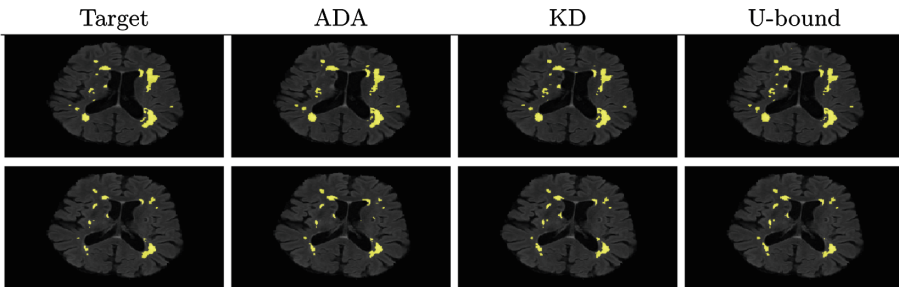

Fig. 1. Illustration of the segmentation's obtained with different methods trained on the Utrecht dataset and tested on the Amsterdam dataset. The top and bottom row illustrate segmentations on two different subjects.

5   Discussion

The main objective of this paper was to present domain adaptation from a semi-supervised learning perspective. We have evaluated a modified knowledge distillation approach and compared it to the popular adversarial approach under different clinical scenarios. Overall, the knowledge distillation approach gave better results and is relatively simpler to design when compared to the more architecture-dependent adversarial approaches. Adversarial approaches require extensive tuning of DNN architectures, especially for the discriminator, in order to achieve reasonable performances. In contrast, KD only involves choosing the temperature parameter which can be chosen only based on the performances on the source domain. One of the interesting outcomes is the inferior performance of KD on domain adaptation in scenario, Utrecht clinic to Singapore clinic. One of the reasons may be attributed to not just scanner differences but also differences in demographics. This may have led to an inferior teacher performance that the student network relies on. To verify this, we used the improved network from domain adaptation using ADA as a teacher and then trained a student based on it. We observed that the mean dice overlap improved from $0.65 \rightarrow 0.69$.

In future work, we will consider combining the adversarial approaches with knowledge distillation to improve the generalisability of DNNs across domains without the need for large annotated datasets.

References

1. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems, pp. 3239–3250 (2018)
2. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
3. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643 (2015)
4. Hoffman, J., Rodner, E., Donahue, J., Darrell, T., Saenko, K.: Efficient learning of domain-invariant image representations. arXiv preprint arXiv:1301.3224 (2013)
5. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 476–484. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_54
6. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176 (2017)
7. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 443–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_35

8. Hoffman, J., et al.: Cycada: cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
9. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
10. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2827–2836 (2016)
11. Huang, R., Noble, J.A., Namburete, A.I.L.: Omni-supervised learning: scaling up to large unlabelled medical datasets. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 572–580. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_65