Unsupervised domain adaptation for medical imaging segmentation with self-ensembling

Christian S. Perone
NeuroPoly Lab
Institute of Biomedical Engineering
Polytechnique Montreal
Montreal, QC, Canada.

Pedro L. Ballester
Machine Intelligence and Robotics Research Group
School of Technology
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, RS, Brazil

Rodrigo C. Barros
Machine Intelligence and Robotics Research Group
School of Technology
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, RS, Brazil

Julien Cohen-Adad
NeuroPoly Lab
Institute of Biomedical Engineering
Polytechnique Montreal
Functional Neuroimaging Unit
Universite de Montreal
Montreal, QC, Canada.

Abstract

Recent deep learning methods for the medical imaging domain have reached state-of-the-art results on several tasks. Those models, however, when trained to reduce the empirical risk on a single domain, fail to generalize to other domains. This is a very common scenario in medical imaging due to the variability of image quality across hospitals and anatomical structures, even for the same imaging modality. In this work, we extend the method of unsupervised domain adaptation using self-ensembling to segmentation tasks and evaluate it on a realistic small data regime using a publicly available MRI dataset. We show evidence that self-ensembling can improve the generalization of the models even when using a small amount of unlabeled data.

1 Introduction

In the past few years, the research community has witnessed the fast developmental pace of deep learning [5] methods for data analysis, establishing an important scientific milestone. Deep neural networks are a paradigm shift from traditional machine learning approaches. While the latter rely on hand-crafted feature engineering, deep neural networks are capable of automatically learning robust hierarchical features, in what is known as *representation learning*.

Due to its popularity and excellent results in many domains, deep learning attracted a lot of attention from the medical imaging community [6]. However, there are still several challenges that need to be properly addressed. For instance, one of the most well-known problems is the high sample complexity, or how much data deep learning requires to accurately learn and perform well on unseen images, which is linked to the concepts of model complexity and generalization, an active research topic in learning theory [7]. The large amount of required data to train deep neural networks can be partially mitigated with techniques such as transfer learning [12, 13]. However, transfer learning is problematic in medical imaging because large datasets are usually required to models take benefit from the inductive transfer process.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Yet, another challenge when deploying deep learning models for medical imaging analysis – and perhaps one of the most difficult to solve – is the so-called *data distribution shift*: the variability inherent to the different imaging protocols (sequence parameters, sites, vendors, scanner model) can result in significantly different data distributions. A concrete example can be found in magnetic resonance imaging (MRI), where the same machine vendor using the same protocol for the same subject can produce different voxel intensities (e.g., variability could be caused by slightly different positioning in the scanner, subject motion and/or different protocol).

Empirical risk minimization (ERM) is the statistical learning principle behind many machine learning methods, and it offers good learning guarantees and bounds if its assumptions hold, such as the fact that the train and test datasets come from the same domain. However, this assumption is usually broken on real scenarios. Although this distributional shift is very common in medical imaging, the problem is surprisingly ignored during the design of many different challenges in the field. It is very common to have the same domain data (same machine, protocol, etc.) on both training and test sets. However, this validation scenario does not represent the reality and in many cases produce over-optimistic evaluation results. The name given to learning a classifier model or any other predictor with a shift between the training and the target/test distributions is known as "domain adaptation" (DA). In this work, we expand a previously-developed method [1] for DA and apply it for a segmentation task, which is the most addressed task in medical imaging [6].

The original contribution of this paper is the extension of the unsupervised domain adaptation method using self-ensembling for the semantic segmentation task. To the best of our knowledge, this is the first time this method is used for semantic segmentation. We perform an extensive evaluation and ablation experiments on a realistic small data regime dataset from the MRI domain.

2 Related Work

In Ganin et al. [2], the authors used adversarial training to devise a method that enforces the network to learn domain-invariant features. This work was later extended to segmentation tasks by Kamnitsas et al. [3] in the medical imaging domain. Recently, state-of-the-art techniques for semi-supervised learning using temporal ensembling were introduced by Laine and Aila [4] and extended to Mean Teachers [11]. Mean Teachers were then adapted to segmentation tasks in the medical imaging domain [8]. In French et al. [1] they extended the temporal ensembling method for domain adaptation on classification tasks. In this work we extend the work done by French et al. [1] to segmentation tasks and evaluate it on a realistic small data regime from the medical imaging domain.

3 Method

Our method is based on the work by French et al. [1] using the Mean Teacher [11] variant. An overview of our method is described in Figure 1. We extended the method for segmentation tasks by using the Dice loss for the segmentation task and the mean squared error (MSE) loss for consistency between student and teacher models.

One of the main challenges of adapting the method for segmentation tasks is the misalignment that is caused by data augmentation on the prediction of both student and teacher models. To overcome this issue we employed the same delayed approach described in Perone and Cohen-Adad [8] to align the predictions from student and teacher before the consistency phase. This was possible because the back-propagation happens only for the student model, so there are no differentiability requirements on any operation between the teacher prediction and the consistency loss computation.

Our technique is robust enough to work with any model architecture because it decouples the domain adaptation component from the model choice. Due to this flexibility, we used one of the most common model architectures for medical imaging, the U-Net [10], which was kept the same both for traditional supervised baseline and for the unsupervised domain adaptation scenario.

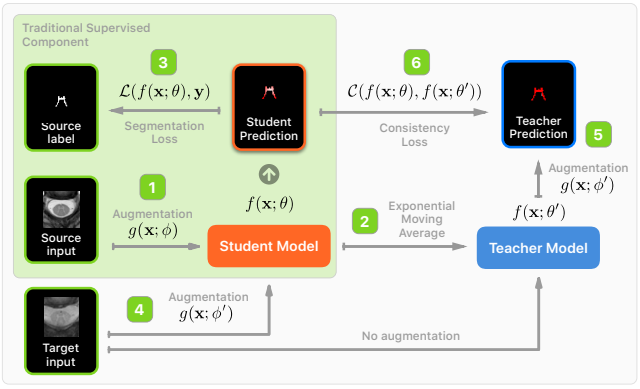

Figure 1: Overview of the proposed method. The green panel represents the traditional supervision signal. **(1)** The source domain input data is augmented by the $g(x;\phi)$ transformation and feed into the student. **(2)** The teacher parameters are updated with an exponential moving average (EMA) from the student weights. **(3)** The traditional segmentation loss. **(4)** The input unlabeled data from the target domain is transformed with $g(x;\phi')$ before the student's forward pass (note different parameters ϕ'). **(5)** The teacher predictions are augmented by the $g(x;\phi')$ transformation. **(6)** The consistency loss. This consistency enforces the consistency between student and teacher predictions.

4 Experiments

4.1 Dataset

We used the Spinal Cord Gray Matter (SCGM) challenge dataset [9]. The SCGM is a multi-center, multi-vendor and publicly-available[1] MRI data collection that is comprised of 80 healthy subjects with 20 subjects from each center. Due to the fact that the SCGM dataset contains data from all 4 centers both in training as well as in the test, we used a non-standard split of the data in order to evaluate our technique on a domain adaptation scenario where the domain present in the test set did not contain contamination from the training data domain. Therefore, we used data from Centers 1 and 2 as the training set, Center 3 as the validation set and Center 4 as the test set.

4.2 Results and conclusions

Table 1 shows results for the unsupervised domain adaptation technique and the traditional supervised technique, with a clear improvement for the domain adaptation technique.

Given that the improvement can also be due to the introduction of the exponential moving average (EMA) alone (by averaging and smoothing the SGD trajectory), to demonstrate that the improvement is specific to the added unlabeled data and not only from the EMA, we performed an ablation experiment by leaving the EMA active and setting the consistency weight to zero. That way we were able to evaluate the impact of the EMA without taking into consideration the unlabeled data. Results of this experiment are described in Table 2, confirming that EMA alone is not enough to explain the improvements found using unlabeled data.

Evaluation	Adaptation	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 3	Baseline	82.81 ± 0.33	71.05 ± 0.36	**90.61** ± 0.63	77.09 ± 0.34	99.86 ± 0.0	2.14 ± 0.02
	Center 3	**84.72** ± 0.18	**73.67** ± 0.28	87.43 ± 1.90	**83.17** ± 1.62	**99.91** ± 0.01	**2.01** ± 0.03
	Center 4	84.45 ± 0.14	73.30 ± 0.19	87.13 ± 1.77	82.92 ± 1.76	**99.91** ± 0.01	2.02 ± 0.03
Center 4	Baseline	69.41 ± 0.27	53.89 ± 0.31	**97.22** ± 0.11	54.95 ± 0.35	99.70 ± 0.00	2.50 ± 0.01
	Center 3	73.27 ± 1.29	58.50 ± 1.57	94.92 ± 1.48	60.93 ± 2.51	99.77 ± 0.03	2.36 ± 0.06
	Center 4	**74.67** ± 1.03	**60.22** ± 1.24	93.33 ± 1.96	**63.62** ± 2.42	**99.80** ± 0.02	**2.29** ± 0.05

Table 1: Evaluation results on validation (Center 3) and test (Center 4) sets. The evaluation and adaptation columns represent, respectively, the centers where testing and adaptation data were collected. The numerical results show the mean and standard deviation over 10 independent runs. Highlighted values represent the best performance metric at each center. All experiments were trained in both centers 1 and 2 simultaneously. mIoU represents the mean Intersection over Union. Other metrics are self-explanatory.

[1]The dataset is available at http://cmictig.cs.ucl.ac.uk/niftyweb/program.php?p=CHALLENGE

Evaluation	Version	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 3	Baseline	83.06	71.36	90.98	77.24	99.86	2.13
	EMA	83.09	71.40	90.97	77.30	99.86	2.13
Center 4	Baseline	69.41	53.90	97.20	54.98	99.70	2.48
	EMA	69.50	54.00	97.19	55.09	99.71	2.48

Table 2: Results of the ablation experiment where the baseline model was trained and compared against its exponential moving average (EMA) model. All experiments were trained in both center 1 and 2 simultaneously. Center 3 is the validation set and Center 4 is the test set.

Acknowledgments

Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [950-230815], the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Fonds de Recherche du Québec - Nature et Technologies [2015-PR-182754], the Natural Sciences and Engineering Research Council of Canada [435897-2013], the Canada First Research Excellence Fund (IVADO and TransMedTech) and the Quebec BioImaging Network [5886]. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001.

References

[1] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[3] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.

[4] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[5] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Lecun Y., Bengio Y., and Hinton G. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

[6] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017. ISSN 13618423. doi: 10.1016/j.media.2017.07.005.

[7] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[8] Christian S. Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. *DLMIA MICCAI*, pages 1–8, sep 2018. doi: 10.1007/978-3-030-00889-5_2.

[9] Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N. Conrad, Esha Datta, Gergely Dávid, Benjamin De Leener, Sara M. Dupont, Patrick Freund, Claudia A.M. Gandini Wheeler-Kingshott, Francesco Grussu, Roland Henry, Bennett A. Landman, Emil Ljungberg, Bailey Lyttle, Sebastien

Ourselin, Nico Papinutto, Salvatore Saporito, Regina Schlaeger, Seth A. Smith, Paul Summers, Roger Tam, Marios C. Yiannakas, Alyssa Zhu, and Julien Cohen-Adad. Spinal cord grey matter segmentation challenge. *NeuroImage*, 152:312–329, 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2017.03.010.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. pages 1–8, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28.

[11] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[12] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems 27 (Proceedings of NIPS)*, 27:1–9, 2014. ISSN 10495258.

[13] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.