Deep Semi-supervised Segmentation with Weight-Averaged Consistency Targets

Christian S. Perone[1] and Julien Cohen-Adad[1,2(✉)]

[1] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal,
Montreal, QC, Canada
jcohen@polymtl.ca
[2] Functional Neuroimaging Unit, CRIUGM, Universite de Montreal, Montreal,
QC, Canada

Abstract. Recently proposed techniques for semi-supervised learning such as Temporal Ensembling and Mean Teacher have achieved state-of-the-art results in many important classification benchmarks. In this work, we expand the Mean Teacher approach to segmentation tasks and show that it can bring important improvements in a realistic small data regime using a publicly available multi-center dataset from the Magnetic Resonance Imaging (MRI) domain. We also devise a method to solve the problems that arise when using traditional data augmentation strategies for segmentation tasks on our new training scheme.

1 Introduction

In the past few years, we witnessed a large growth in the development of Deep Learning techniques, that surpassed human-level performance on some important tasks [1], including health domain applications [2]. A recent survey [3] that examined more than 300 papers using Deep Learning techniques in medical imaging analysis, made it clear that Deep Learning is now pervasive across the entire field. In [3], they also found that Convolutional Neural Networks (CNNs) were more prevalent in the medical imaging analysis, with end-to-end trained CNNs becoming the preferred approach.

It is also evident that Deep Learning poses unique challenges, such as the large amount of data requirement, which can be partially mitigated by using transfer learning [4] or domain adaptation approaches [5], especially in the natural imaging domain. However, in medical imaging domain, not only the image acquisition is expensive but also data annotations, that usually requires a very time-consuming dedication of experts. Besides that, other challenges are still present in the medical imaging field, such as privacy and regulations/ethical concerns, which are also an important factor impacting the data availability.

According to [3], in certain domains, the main challenge is usually not the availability of the image data itself, but the lack of relevant annotations/labeling

for these images. Traditionally, systems like Picture Archiving and Communication System (PACS) [3], used in the routine of most western hospitals, store free-text reports, and turning this textual information into accurate or structured labels can be quite challenging. Therefore, the development of techniques that could take advantage of the vast amount of unlabeled data is paramount for advancing the current state of practical applications in medical imaging.

Semi-supervised learning is a class of learning algorithms that can take leverage not only of labeled samples but also from unlabeled samples. Semi-supervised learning is halfway between supervised learning and unsupervised learning [6], where the algorithm uses limited supervision, usually only from a few samples of a dataset together with a larger amount of unlabeled data.

In this work, we propose a simple deep semi-supervised learning approach for segmentation that can be efficiently implemented. Our technique is robust enough to be incorporated in most traditional segmentation architectures since it decouples the semi-supervised training from the architectural choices. We show experimentally on a public Magnetic Resonance Imaging (MRI) dataset that this technique can take advantage of unlabeled data and provide improvements even in a realistic scenario of small data regime, a common reality in medical imaging.

2   Semi-supervised Segmentation Using Mean Teacher

Given that the classification cost for unlabeled samples is undefined in supervised learning, adding unlabeled samples into the training procedure can be quite challenging. Traditionally, there is a dataset $\mathbf{X} = (x_i)_{i \in [n]}$ that can be divided into two disjoint sets: the samples $\mathbf{X}_l = (x_1, \ldots, x_l)$ that contains the labels $\mathbf{Y}_l = (y_1, \ldots, y_l)$, and the samples $\mathbf{X}_u = (x_{l+1}, \ldots, x_{l+u})$ where the labels are unknown. However, if the knowledge available in $p(x)$ that we can get from the unlabeled data also contains information that is useful for the inference problem of $p(y|x)$, then it is evident that semi-supervised learning can improve upon supervised learning [6].

Many techniques were developed in the past for semi-supervised learning, usually creating surrogate classes as in [7], adding entropy regularization as in [8] or using Generative Adversarial Networks (GANs) [9]. More recently, other ideas also led to the development of techniques that added perturbations and extra reconstruction costs in the intermediate representations [10] of the network, yielding excellent results. A very successful method called Temporal Ensembling [11] was also recently introduced, where the authors explored the idea of a temporal ensembling network for semi-supervised learning where the predictions of multiple previous network evaluations were aggregated using an exponential moving average (EMA) with a penalization term for the predictions that were inconsistent with this target, achieving state-of-the-art results in several semi-supervised learning benchmarks.

In [12], the authors expanded the Temporal Ensembling method by averaging the model weights instead of the label predictions by using Polyak averaging

[13]. The method described in [12] is a student/teacher model, where the student model architecture is replicated into the teacher model, which in turn, get its weights updated as the exponential moving average of the student weights according to:

$$\theta'_t = \alpha\theta'_{t-1} + (1 - \alpha)\theta_t \tag{1}$$

where α is a smoothing hyperparameter, t is the training step and θ are the model weights. The goal of the student is to learn through a composite loss function with two terms: one for the traditional classification loss and another to enforce the consistency of its predictions with the teacher model. Both the student and teacher models evaluate the input data by applying noise that can come from Dropout, random affine transformations, added Gaussian noise, among others.

In this work, we extend the mean teacher technique [12] to semi-supervised segmentation. To the best of our knowledge, this is the first time that this semi-supervised method was extended for segmentation tasks. Our changes to extend the mean teacher [12] technique for segmentation are simple: we use different loss functions both for the task and consistency and also propose a new method for solving the augmentation issues that arises from this technique when used for segmentation. For the consistency loss, we use a pixel-wise binary cross-entropy, formulated as

$$\mathcal{C}(\theta) = \mathbb{E}_{x \in \mathbf{X}}\left[-y\log(p) + (1 - y)\ log(1 - p)\right], \tag{2}$$

where $p \in [0, 1]$ is the output (after sigmoid activation) of the student model $f(x; \theta)$ and $y \in [0, 1]$ is the output prediction for the same sample from the teacher model $f(x; \theta')$, where θ and θ' are student and teacher model parameters respectively. The consistency loss can be seen as a pixel-wise knowledge distillation [14] from the teacher model. It is important to note that both labeled samples from $\mathbf{X}_l$ and unlabeled samples from $\mathbf{X}_u$ contribute for the consistency loss $\mathcal{C}(\theta)$ calculation. We used binary cross-entropy, instead of the mean squared error (MSE) used by [12] because the binary cross-entropy provided an improved model performance for the segmentation task. We also experimented with confidence thresholding as in [15] on the teacher predictions, however, it didn't improve the results.

For the segmentation task, we employed a surrogate loss for the Dice Similarity Coefficient, called the Dice loss, which is insensitive to imbalance and was also employed by [16] on the same segmentation task domain we experiment in this paper. The Dice Loss, computed per mini-batch, is formulated as

$$L(\theta) = -\frac{2\sum_i p_i y_i}{\sum_i p_i + \sum_i y_i}, \tag{3}$$

where $p_i \in [0, 1]$ is the ith output (after sigmoid non-linearity) and $y_i \in \{0, 1\}$ is the corresponding ground truth. For the segmentation loss, only labeled samples from $\mathbf{X}_l$ contribute for the $\mathcal{L}(\theta)$ calculation. As in [12], the total loss used is the weighted sum of both segmentation and consistency losses. An overview detailing the components of the method can be seen in the Fig. 1, while a description of the training algorithm is described in the Algorithm 1.
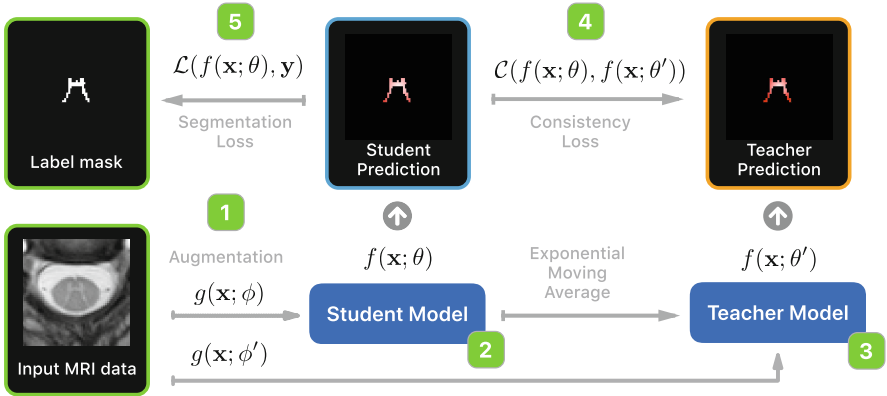
Fig. 1. An overview with the components of the proposed method based on the mean teacher technique. **(1)** A data augmentation procedure $g(x; \phi)$ is used to perturb the input data (in our case, a MRI axial slice), where ϕ is the data augmentation parameter (i.e. $\mathcal{N}(0, \phi)$ for a Gaussian noise), note that different augmentation parameters are used for student and teacher models. **(2)** The student model. **(3)** The teacher model that is updated with an exponential moving average (EMA) from the student weights. **(4)** The consistency loss used to train the student model. This consistency will enforce the consistency between student predictions on both labeled and unlabelled data according to the teacher predictions. **(5)** The traditional segmentation loss, where the supervision signal is provided to the student model for the labeled samples.

2.1   Segmentation Data Augmentation

In segmentation tasks, data augmentation is very important, especially in the medical imaging domain where data availability is limited, variability is high and translational equivariance is desirable. Traditional augmentation methods such as affine transformations (rotation, translation, etc.) that change the spatial content of the input data, as opposed to pixel-wise additive noise, for example, are also applied with the exact same parameters on the label to spatially align input and ground truth, both subject to a pixel-wise loss. This methodology, however, is unfeasible in the mean teacher training scheme. If two different augmentations (one for the student and another for the teacher) causes spatial misalignment, the spatial content between student and teacher predictions will not match during the pixel-wise consistency loss.

   To avoid the misalignment during the consistency loss, such transformations can be applied with the same parametrization both to the student and teacher model inputs. However, this wouldn't take advantage of the stronger invariance to transformations that can be introduced through the consistency loss. For that reason, we developed a solution that applies the transformations in the teacher in a delayed fashion. Our proposed method is based on the application of the same augmentation procedure $g(x; \phi)$ before the model forward pass only for the student model, and then after model forward pass in the teacher model predictions, making thus both prediction maps aligned for the consistency loss evaluation,

Algorithm 1. Semi-supervised segmentation algorithm.

Require: x_i = training samples
Require: y_i = labels for the labeled inputs $i \in \mathbf{Y}_l$
Require: t = global step (initialized with zero)
Require: $w(t)$ = consistency weight ramp-up function
Require: $f_\theta(\cdot)$ = neural network model with parameters θ
Require: $g_\phi(\cdot)$ = stochastic input augmentation procedure with parameters ϕ
  for k in $[1, num\_epochs]$ **do**
    for each minibatch B **do**
      $z_{i \in B} \leftarrow f_\theta(g_\phi(x_{i \in B}))$ ▷ evaluate augmented inputs with student model
      $\tilde{z}_{i \in B} \leftarrow f_{\theta'}(g_{\phi'}(x_{i \in B}))$ ▷ teacher model evaluation w/ different perturbations
      $loss \leftarrow \mathcal{L}(z, y) + w(t) \frac{1}{|B|} \sum_{i \in B} \mathcal{C}(z_i, \tilde{z}_i)$ ▷ supervised and unsupervised loss components
      update θ using, e.g., ADAM   ▷ update student model parameters
      $t \leftarrow t + 1$ ▷ increment the global step counter
      $\theta'_t \leftarrow \alpha\theta'_{t-1} + (1 - \alpha)\theta_t$ ▷ update teacher model parameters with using EMA
    end for
  end for

while still taking leverage of introducing a much stronger invariance to the augmentation between student and teacher models. This is possible because we do backpropagation of the gradients only for the student model parameters.

3 Experiments

3.1 MRI Spinal Cord Gray Matter Segmentation

In this work, in order to evaluate our technique on a realistic scenario, we use the publicly available multi-center Magnetic Resonance Imaging (MRI) Spinal Cord Gray Matter Segmentation dataset from [17].

Dataset. The dataset is comprised of 80 healthy subjects (20 subjects from each center) and obtained using different scanning parameters and also multiple MRI systems. The voxel resolution of the dataset ranges from $0.25 \times 0.25 \times 2.5$ mm up to $0.5 \times 0.5 \times 5.0$ mm. A sample of one subject axial slice image can be seen in Fig. 1. We split the dataset in a realistic small data regime: only 8 subjects are used as training samples, resulting in 86 axial training slices. We used 8 subjects for validation, resulting in 90 axial slices. For the unlabeled set we used 40 subjects, resulting in 613 axial slices and for the test set we used 12 subjects, resulting in 137 slices. All samples were resampled to a common space of 0.25×0.25 mm.

Network Architecture. To evaluate our technique, we used a very simple U-Net [18] architecture with 15 layers, Batch Normalization, Dropout and ReLU activations. U-Nets are very common in medical imaging domain, hence the architectural choice for the experiment. We also used a 2D slice-wise training procedure with axial slices.

Training Procedure. For the supervised-only baseline, we used Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, mini-batch size of 8, dropout rate of 0.5, Batch Normalization momentum of 0.9 and L2 penalty of $\lambda = 0.0008$. For the data augmentation, we used rotation angle between -4.5 and 4.5 and pixel-wise additive Gaussian noise sampled from $\mathcal{N}(0, 0.01)$. We used a learning rate $\eta = 0.0006$ given the small mini-batch size, also subject to a initial ramp-up of 50 epochs and subject to a cosine annealing decay as used by [12]. We trained the model for 1600 epochs.

For the semi-supervised experiment, we used the same parameters of the aforementioned supervised-only baseline, except for the L2 penalty of $\lambda = 0.0006$. We used an EMA $\alpha = 0.99$ during the first 50 epochs, later we change it to $\alpha = 0.999$. We also employed a consistency weight factor of 2.9 subject to a ramp-up in the first 100 epochs. We trained the model for 350 epochs.

Results. As we can see in Table 1, our technique not only improved the results on 5/6 evaluated metrics but also reduced the variance, showing a better regularized model in terms of precision/recall balance. The model also showed a very good improvement on overlapping metrics such as Dice and mean intersection over union (mIoU). Given that we exhausted the challenge dataset [17] to obtain the unlabeled samples, a comparison with [16] was unfeasible given different dataset splits. We leave this work for further explorations given that incorporating extra external data would also mix domain adaptation issues into the evaluation.

Table 1. Result comparison for the Spinal Cord Gray Matter segmentation challenge using our semi-supervised method and a pure supervised baseline. Results are 10 runs average with standard deviation in parenthesis where bold font represents the best result. Dice is the Dice Similarity Coefficient and mIoU is the mean intersection over union. Other metrics are self-explanatory.

	Dice	mIoU	Accuracy	Precision	Recall	Specificity
Supervised	67.915	53.679	99.745	57.948	**92.495**	99.775
	(0.313)	(0.327)	(0.005)	(0.788)	**(0.907)**	(0.010)
Semi-supervised	**70.209**	**55.509**	**99.792**	**64.732**	86.112	**99.846**
	(0.229)	**(0.253)**	**(0.003)**	**(0.773)**	(0.936)	**(0.006)**

4   Related Work

Only a few works were developed in the context of semi-supervised segmentation, especially in the field of medical imaging. Only recently, a U-Net was used as auxiliary embedding in [19], however, with focus on domain adaptation and using a private dataset.

In [20], they use a Generative Adversarial Networks (GAN) for the semi-supervised segmentation of natural images, however, they employ unrealistic

dataset sizes when compared to the medical imaging domain datasets, along with ImageNet pre-trained networks.

In [21] they propose a technique using adversarial training, but they focus on the knowledge transfer between natural images with pixel-level annotation and weakly-labeled images with image-level information.

5   Conclusion

In this work we extended the semi-supervised mean teacher approach for segmentation tasks, showing that even on a realistic small data regime, this technique can provide major improvements if unlabeled data is available. We also devised a way to maintain the traditional data augmentation procedures while still taking advantage of the teacher/student regularization. The proposed technique can be used with any other Deep Learning architecture since it decouples the semi-supervised training procedure from the architectural choices.

It is evident from these results that future explorations of this technique can improve the results even further, given that even with a small amount of unlabeled samples, we showed that the technique was able to provide significant improvements.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, 11–18 December, pp. 1026–1034 (2016)
2. Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y.: Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint (2017)
3. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
4. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems (Proceedings of NIPS), vol. 27, pp. 1–9 (2014)
5. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**, 1–35 (2016)
6. Olivier, C., Schölkopf, B., Zien, A.: Semi-supervised learning. Interdiscip. Sci. Comput. Life Sci. **1**(2), 524 (2006)

7. Lee, D.H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: ICML 2013 Workshop: Challenges in Representation Learning, pp. 1–6 (2013)

8. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems - NIPS 2004, vol. 17, pp. 529–536 (2004)

9. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs (2016)

10. Rasmus, A., Valpola, H., Berglund, M.: Semi-supervised learning with ladder network, pp. 1–17 (2015). arXiv

11. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning (2016)

12. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results (2017)

13. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control. Optim. **30**(4), 838–855 (1992)

14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network, March 2015

15. French, G., Mackiewicz, M., Fisher, M.: Self-ensembling for domain adaptation, pp. 1–15 (2017)

16. Perone, C.S., Calabrese, E., Cohen-Adad, J.: Spinal cord gray matter segmentation using deep dilated convolutions, October 2017

17. Prados, F., et al.: Spinal cord grey matter segmentation challenge. NeuroImage **152**, 312–329 (2017)

18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

19. Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 311–319. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_36

20. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5689–5697. IEEE, October 2017

21. Xiao, H., Wei, Y., Liu, Y., Zhang, M., Feng, J.: Transferable semi-supervised semantic segmentation (2017)