Teacher-student semi-supervised approach for medical image segmentation

Maria Baldeon Calisto[1[0000−0001−9379−8151]]

Departamento de IngenierÃŋa Industrial and Instituto de InnovaciÃşn en Productividad y LogÃŋstica CATENA-USFQ, Universidad San Francisco de Quito, Diego de Robles s/n y VÃŋa InteroceÃąnica, Quito, Ecuador 170901
mbaldeonc@usfq.edu.ec,

Abstract. Accurate segmentation of anatomical structures is a critical step for medical image analysis. Deep learning architectures have become the state-of-the-art models for automatic medical image segmentation. However, these models require an extensive labelled dataset to achieve a high performance. Given that obtaining annotated medical datasets is very expensive, in this work we present a two-phase teacher-student approach for semi-supervised learning. In phase 1, a 5-network U-Net ensemble, denominated the teacher, is trained using the labelled dataset. In phase 2, a student U-Net network is trained with the labelled dataset and the unlabelled dataset with the pseudo-labels produced with the teacher network. The student network is then used inference of the testing images. The proposed approach is evaluated on the task of abdominal segmentation from the FLARE2022 challenge, achieving a mean 0.53 dice, 0.57 NSD, and 44.97 prediction time on the validations set.

Keywords: Semi-supervised learning · Image Segmentation · Medical Image Analysis.

1 Introduction

Accurate segmentation of anatomical structures is a critical step for medical image analysis. Deep learning models have become the de-facto techniques for segmentation tasks given its state-of-the-art performance in various medical datasets. However, without an extensive labelled dataset, neural networks can overfit the training data and perform poorly in unseen data points. In the case of medical image segmentation, this is an important limitation because although medical imagining datasets have been growing in size, annotating segmentation masks is an expensive and laborious process that requires of an experienced radiologists. Therefore it has become necessary to develop models that leverage unlabeled data information to aid the learning process.

A promising research direction is semi-supervised learning (SSL). SSL models aim to utilize information from unlabelled data to produce predictions that achieve a higher performance than if trained solely with labelled data [11]. Recently, important semi-supervised deep learning models have been proposed for

medical image segmentation. [6] developed a dual-task network that predicts the pixel-wise segmentation map and level set function of the input image. The implemented loss function combines a supervised learning loss with an unsupervised dual-task-consistency loss function. [2] proposed a multi-task attention-based SSL model that combines an autoencoder with a U-Net-like network. The autoencoder is trained to reconstruct synthetic segmentation labels that encourages the segmentation model to learn discriminative latent representations from the unlabelled images. One limitation of previous approaches is that the models are tested in datasets where the number of classes is small, so when the numbers of classes increases, the complexity and size of the training framework rises importantly.

In this work, we propose a two-phase teacher-student semi-supervised training approach. In phase 1, a three-network "teacher" ensemble is trained in a supervised manner using the labelled dataset. In phase 2, a "student" network is trained with the labelled images in a supervised manner, and the unlabelled images using the pseudo-labels provided by the teacher network. The model is tested on the FLARE2022 challenge dataset, that aims to segment 13 abdominal organs. Our model achieves a mean 0.53 dice, 0.57 NSD, and 44.97 prediction time on the validations set. Our experiments demonstrate that using the teacher-student approach increases a 2% the dice metric over using only the labelled dataset.

2 Method

The proposed method is composed of two phases as displayed in Figure 1. In phase one, three 2D U-Net [9] models are trained in a supervised manner with the labelled training set using a 5-fold cross validation division scheme. The teacher network is formed by uniting the networks through soft voting. In phase two, a student 2D U-Net is trained in a semi-supervised manner with the labelled and unlabelled images with pseudo-labels provided by the teacher ensemble. Details of each phase are provided next.

2.1 Phase One

The training dataset is divided into 5 folds, by assigning 80% of the observations for training and 20% observations for validation. A deeply supervised 2D U-Net, as presented in Figure 2, is trained on each of the folds using the training protocols described in the following section. The 2D U-Net is composed of five down-sampling modules and four up-sampling modules. The modules are comprised of two convolutional blocks, each convolutional block having a 3×3 convolutional layer, batch normalization layer, and ReLU activation function. The last and second-last up-sampling modules are followed a 1×1 convolutional layer with a softmax activation function to produce the predicted segmentation. The objective function being minimized during training is a linear combination

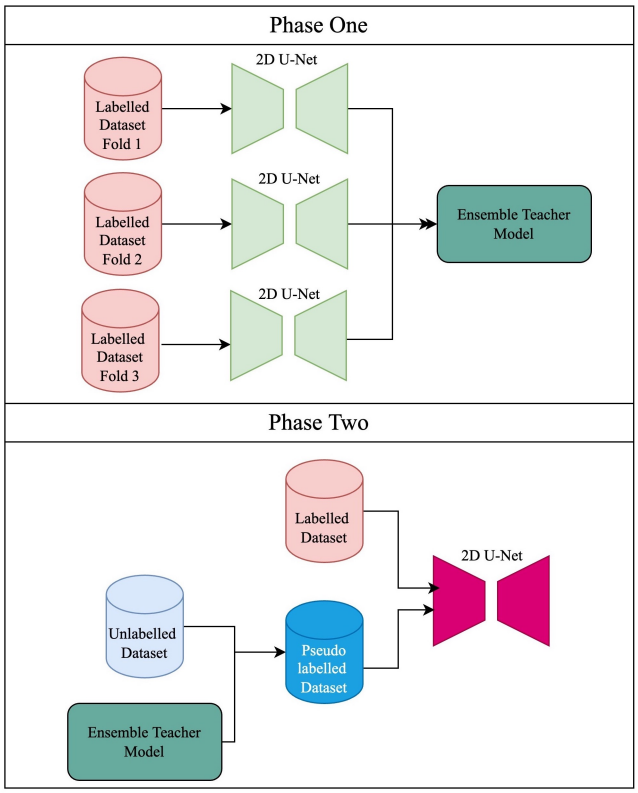

Fig. 1. Two-Phase Approach for semi-supervised learning. In Phase 1 a teacher ensemble is trained in a supervised manner. In Phase 2 a student network is trained in an semi-supervised manner using pseudo labels from the teacher.

of the soft dice loss and cross entropy loss as presented in Eq. 1 as it has shown to provide robust results in various medical image segmentation tasks [7].

$$\mathcal{L}_{seg} = \beta \sum_{c} 1 - \frac{2\sum_i \widehat{y}_{ic} y_{ic}}{\sum_i \widehat{y}_{ic} + \sum_i y_{ic}} + (1-\beta) \sum_{c} \sum_{i} (y_{ic} log(\widehat{y}_{ic}) + (1-y_{ic}) log(1-\widehat{y}_{ic})) \tag{1}$$

where y_{ic} is the ground-truth label for pixel i in class c, and $\widehat{y}_{ic}$ the corresponding predicted probability. β is a weight parameter for the dice loss, which we set to 0.65. As previously mentioned, a deep supervised layer with an auxiliary segmentation loss [**?**] is located in the second-last up-sampling block to aid the model to learn rich hierarchical features. Therefore, the final loss function is comprised of the loss from the main output and the loss from the deep supervised layer with a weight of 0.1.

From the five networks trained, three are selected based on their performance on the challengeÂt's validation set to form an ensemble.

2.2 Phase Two

In phase two, a 2D U-Net architecture (refer to Figure 2) is trained in a semi-supervised manner to segment the medical images. First, the teacher ensemble network formed in phase one is utilized to produce pseudo-labels for the unlabelled images. During a training iteration the student 2D U-Net is trained with a batch of 2D labelled images using the same loss function displayed in Eq. 1, and later with a batch of unlabelled images with the psuedo-labels as ground truth with the loss shown in Eq. 2.

$$\mathcal{L}_{pseudo-seg} = \beta \sum_{c} 1 - \frac{2\sum_i \widehat{y}_{ic} \tilde{y}_{ic}}{\sum_i \widehat{y}_{ic} + \sum_i \tilde{y}_{ic}} + (1-\beta) \sum_{c} \sum_{i} (\tilde{y}_{ic} log(\widehat{y}_{ic}) + (1-\tilde{y}_{ic}) log(1-\widehat{y}_{ic})) + \sum_{i} \sum_{c} ||\widehat{y}_{ic} - \tilde{y}_{ic}|| \tag{2}$$

The resulting 2D U-Net is used for inference on the validation and testing test. This trick also allows the single 2D U-Net to learn all the information of the three-network ensemble, performing even better than the ensemble while the reducing the size to 1/3.

3 Experiments

3.1 Dataset and evaluation measures

The FLARE2022 dataset is curated from more than 20 medical groups under the license permission, including MSD [10], KiTS [4,5], AbdomenCT-1K [8], and TCIA [3]. The training set includes 50 labelled CT scans with pancreas disease and 2000 unlabelled CT scans with liver, kidney, spleen, or pancreas

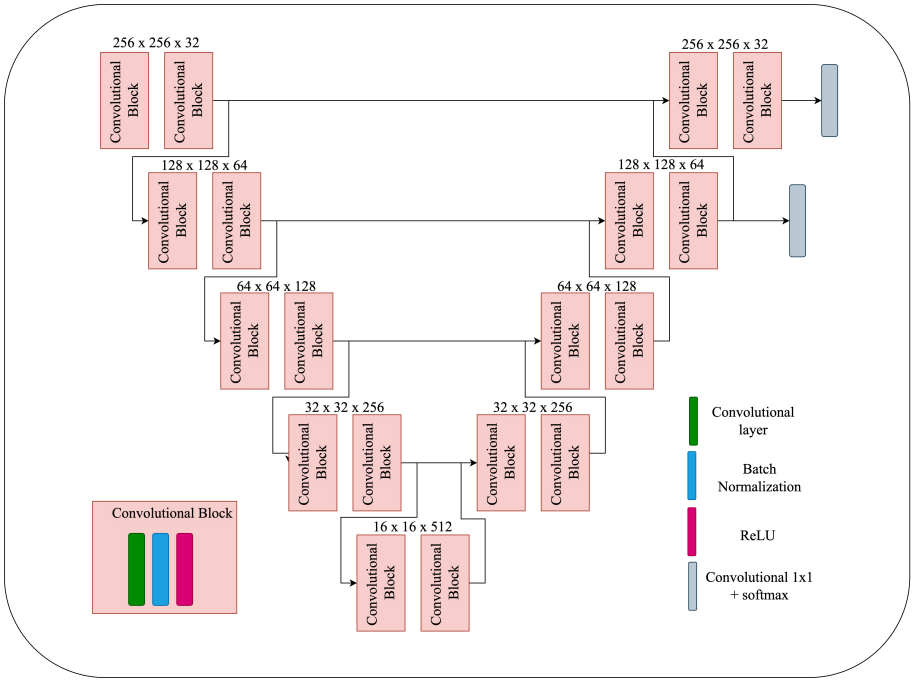

Fig. 2. 2D U-Net implemented to segment the abdominal structures. It is composed of five down-sampling modules and four up-sampling modules. The modules are comprised of two convolutional blocks, each convolutional block having a 3×3 convolutional layer, batch normalization layer, and ReLU activation function.

diseases. The validation set includes 50 CT scans with liver, kidney, spleen, or pancreas diseases. The testing set includes 200 CT scans where 100 cases has liver, kidney, spleen, or pancreas diseases and the other 100 cases has uterine corpus endometrial, urothelial bladder, stomach, sarcomas, or ovarian diseases. All the CT scans only have image information and the center information is not available.

The evaluation measures consist of two accuracy measures: Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), and three running efficiency measures: running time, area under GPU memory-time curve, and area under CPU utilization-time curve. All measures will be used to compute the ranking. Moreover, the GPU memory consumption has a 2 GB tolerance.

3.2 Preprocessing

The images have a heterogeneous voxel spacing and shape. Hence, we first resample all images to have a voxel spacing of $1.5mm \times 1.5mm \times 2.5mm$ and set to a fixed size of $256 \times 256 \times 123$ voxels. Moreover, the pixel intensities are clipped to be inside the 3 standard deviations from the mean and rescaled to a [0 , 1] range.

3.3 Post-processing

No post-processing operations are applied.

3.4 Implementation details

Environment settings The environments and requirements are presented in Table 1.

Table 1. Environments and requirements.

Windows/Ubuntu version	Ubuntu 18.04
CPU	Intel Xeon E5-2698
RAM	256 GB
GPU (number and type)	Four Nvidia V100 32G
CUDA version	10.1.243
Programming language	Python 3.9
Deep learning framework	Pytorch (Torch 1.11, torchvision 0.12.0)
Specific dependencies	SimpleITK, nibabel, numpy, albumentation

Training protocols The training protocols for phase one and phase two are presented in Table 2 and Table 3 respectively. On both phases we implement data augmentation on the fly for the labelled dataset using the albumentations

library [1]. The operations implemented are horizontal flip, vertical flip, random rotation to a maximum of +/- 90 degrees, elastic transformation, grid distortion, and optical distortion.

Table 2. Training protocol phase one.

Network initialization	Kaiming Uniform
Batch size	40
Patch size	256×256
Total epochs	3000
Optimizer	ADAM ($\beta_1 = 0.5$, $\beta_2 = 0.999$)
Initial learning rate (lr)	0.0002
Lr with polynomial decay	
Training time	216 hours
Number of model parameters	6.98M each 2D U-Net network
Number of flops	9.07G each 2D U-Net network
Loss function	Dice loss + Cross-entropy loss

Table 3. Training protocols for phase two

Network initialization	Kaiming Uniform
Batch size	20
Patch size	256×256
Total epochs	200
Optimizer	ADAM ($\beta_1 = 0.5$, $\beta_2 = 0.999$)
Initial learning rate (lr)	1×10^{-5}
Lr with polynomial decay	
Training time	48 hours
Number of model parameters	6.98M
Number of flops	9.07G
Loss function	Dice loss + Cross-entropy loss

4 Results and discussion

4.1 Quantitative results on validation set

The proposed model is tested in the validation set and the evaluation metrics obtained through the challengeÂt's website and displayed in Table 5. We first test the U-Net trained with all the labelled images, which obtained a mean dice

of 0.4815. We also test the three U-Net network ensemble obtained in Phase 1, which achieved a 0.4973 mean dice. Finally, we test the proposed semi-supervised phase 1 and phase 2 approach, which increased in approximately 0.02 the mean dice. Due to computational limitations, we were not able to train with all the unlabelled images and had to selected a subset of 750 images for the implementation of phase 2. In Table **??** the validation scores for each substructure segmented is shown.

Table 4. Evaluation metrics on the validation set

Network	Mean DSC
U-Net (supervised training)	0.4815
Ensemble U-Net (supervised training)	0.4973
Phase 1 + Phase 2	0.5272

Table 5. Evaluation metrics per substructre on the validation set

Substructure	Mean DSC	Mean NSD
Liver	0.74 ± 0.25	0.66 ± 0.24
Right Kidney	0.56 ± 0.38	0.55 ± 0.35
Spleen	0.53 ± 0.35	0.51 ± 0.32
Pancreas	0.49 ± 0.30	0.62 ± 0.30
Aorta	0.67 ± 0.26	0.70 ± 0.25
Inferior Vena Cava	0.59 ± 0.28	0.58 ± 0.28
Right Adrenal Gland	0.46 ± 0.31	0.58 ± 0.36
Left Adrenal Gland	0.43 ± 0.32	0.53 ± 0.37
Gallbladder	0.46 ± 0.39	0.45 ± 0.38
Esophagus	0.37 ± 0.37	0.43 ± 0.42
Stomach	0.56 ± 0.29	0.57 ± 0.25
Duodenum	0.40 ± 0.27	0.62 ± 0.28
Left Kidney	0.58 ±0.37	0.56 ± 0.35

4.2 Qualitative results on validation set

Figure 3) presents examples with good segmentation results and examples with poor segmentation results. The algorithm performs better in the segmentation of the liver, aorta, and inferior vena cava. Meanwhile, it has problems recognizing and segmenting the duodenum and esophagus.

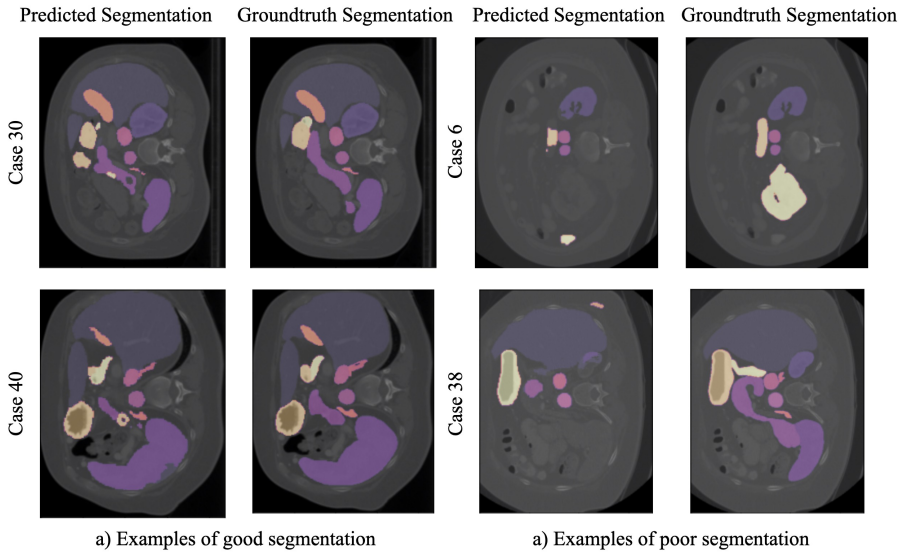

Fig. 3. Examples of good and poor performing segmentation

4.3 Segmentation efficiency results on validation set

The average segmentation efficiency results on the validation set are presented in Table 6.

Table 6. Average segmentation efficiency metrics on the validation set

Time	44.97
Max GPU Memory	1405
AUC GPU Time	49930.08
Max CPU Utilization	95.10
AUC CPU Time	785.13

4.4 Limitation and future work

A big limitation was the computational memory available. The computing infrastructure is shared between various users, so it was impossible to use all the unlabelled images during training. For future work, we will analyze the confidence of the pseudo labels and implement a GAN to force all segmentations to follow a similar distribution.

5 Conclusion

In the present work we propose a two-phase semi-supervised learning approach. In the first phase a three-network "teacher" ensemble is formed by using only the labelled training set. In the second phase, a segmentation network is trained in a semi-supervised scheme using the labelled dataset and unlabelled dataset with pseudo-labels provided by the "teacher" ensemble. Phase two improved in approximately 0.02 the mean dice.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2022 challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers.

References

1. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2) (2020). https://doi.org/10.3390/info11020125, https://www.mdpi.com/2078-2489/11/2/125 6
2. Chen, S., Bortsova, G., García-Uceda Juárez, A., Tulder, G.v., Bruijne, M.d.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–465. Springer (2019) 2
3. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 4
4. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 4
5. Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging. American Society of Clinical Oncology **38**(6), 626–626 (2020) 4
6. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8801–8809 (2021) 2
7. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021) 4
8. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2021.3100536 4

9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015) 2
10. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 4
11. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning **109**(2), 373–440 (2020) 1