A SIMPLE CONSISTENCY REGULARIZATION METHOD FOR SEMI-SUPERVISED ABDOMINAL ORGAN SEGMENTATION

Yanyu Ma[1]

[1] East China University of Science and Technology, No. 130 Meilong Road, Shanghai
chris.y.ma98@gmail.com

Abstract. The scarcity of pixel-level annotation is a prevalent problem in medical image segmentation tasks. In this paper, we applied a novel regularization strategy involving interpolation-based mixing for semi-supervised medical image segmentation. The proposed method is a new consistency regularization strategy that encourages segmentation of interpolation of two unlabeled data to be consistent with the interpolation of segmentation maps of those data. This method represents a specific type of data-adaptive regularization paradigm which aids to minimize the overfitting of labelled data under high confidence values. This method is originally used on ACDC and MMWHS datasets. When applied to the FLARE22 dataset, this method shows excellent performances in inference time and CPU/GPU consumption, yet proves to be insufficient in accuracy. The validation result shows that it can only segment liver with DSC of 0.41, and fails to segment other organs with their DSC under 0.1. It appears that this method with 2D convolution, through simple and efficient, is unsuitable for the FLARE22 dataset and its task. More experiments need to be conducted to confirm whether this method can be used for complex tasks like FLARE22 after major modifications.

Keywords: Semi-supervised Learning, Medical Image Segmentation, Interpolation, Consistency Regularization

1 Introduction

Abdomen organ segmentation has many important clinical applications, such as organ quantification, surgical planning, and disease diagnosis. Supervised medical image segmentation is widely explored problem in computer vision, achieving exponential growth recently. These methods mostly rely upon deep learning, requiring large-scale pixel-wise annotation data. However, manually annotating organs from CT scans is time-consuming and labor-intensive. Semi-Supervised Learning (SSL) is a promising direction in this regard, requiring only a few labelled data and compensating for the large portion of unlabeled data by generating pseudo labels. Recently, SSL-based methods have been widely recognized for their superior performance in medical image segmentation. They not only eliminate the necessity of large-scale

annotations but also produce accurate segmentation results that are very close to those obtained from supervised models.

The FLARE dataset is composed of a small number of labeled case(50) and a large number of unlabeled cases(2000) in the training set, 50 visible cases for validation, and 200 hidden cases for testing. The segmentation targets include 13 organs: liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum. In addition to the typical Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), our evaluation metrics also focus on the inference speed and resources (GPU, CPU) consumption.

As for semi-supervised learning, consistency regularization is a plausible solution, that encourages realistic perturbations of an unlabeled image to produce consistent segmentation maps. For example, Bortsova et al. [1] proposes a transformation consistent segmentation network, capable of exploring the equivariance of elastic perturbations for precise lung X-ray segmentation and Li et al. [2] proposes a semi-supervised segmentation framework using transformation consistency, encouraging consistent predictions of the network-in-training for different perturbation of the same input. Unlike these methods, which rely upon the low-density region assumption, our proposed method chooses perturbation directed towards another unlabeled sample, thereby reducing the necessity of expensive gradient calculation.

On the other hand, interpolation-based regularizes have been achieving state-of-the-art performance in various tasks and across multiple architectures. Recently, this idea has been extended to an unsupervised setting by Berthelot et al. [3] where the authors propose that utilizing realism of the latent space interpolation from autoencoder can improve model learning. Driven by this speculation and the aforementioned success of consistency regularization in SSL methods, in this paper we applied an interpolation-based mixing technique in a semi-supervised setting and its utility in medical image segmentation.

When applied on the FLARE22 dataset, most of the modifications is on preprocess part. And the semi-supervised part and objective function are identical to the original method introduce by the original paper[4].

2 Methodology

2.1 Dataset and Preprocessing

The original file contains 50 labeled and 2000 unlabeled .nii.gz file. The cases are converted into numpy format and sliced into 2d images, because the used method of consistency regularization is based on 2d convolution kernels and 2d interpolation. Besides, due to the high consumption of 3d convolution , I wondered that this 2d network can show higher efficiency.

I noticed that the Z dimensions in this dataset vary significantly, so all files are resampled to [z=64]. And X,Y dimensions are resampled to [256,256]. The resample

function is borrowed directly from nnUnet[5]. After resampling, the converted numpy array shape is [256,256,64] for all data in the datasets.

Then the 3d array is sliced along Z, each file is sliced into 64 slices with shapes of [256,256]. The slices are saved as .h5 format. Data augmentation to the slices is a random combination of flip and rotate, which is identical to the original method applied on ACDC dataset. In the end, all 50 labeled cased are converted to 3,200 slices along with their labels. And the unlabeled cases are converted into 12,8000 slices. But in the practice of training, I only used 1000 unlabeled cases, that to say 6,4000 slices.

2.2 Consistency Regularization with Interpolation

Consistency Regularization has been used in existing literature as a means to enhance the robustness of a network pipeline by enforcing the network against various perturbations on the unlabeled data, to increase the generalizability of these new data points. Among them, the most effective perturbation would be in the adversarial direction – in the direction almost perpendicular to the decision boundary between the positive and negative examples, i.e., the direction in which the network is most liable to misclassify the pixels. However, most existing literature incorporates perturbations that may or may not be in the adversarial direction, and thus, results in loss of generalizability. Some other methods do perturb the input in the adversarial direction but require a large amount of unlabeled data and thus, might not be feasible in the biomedical domain.

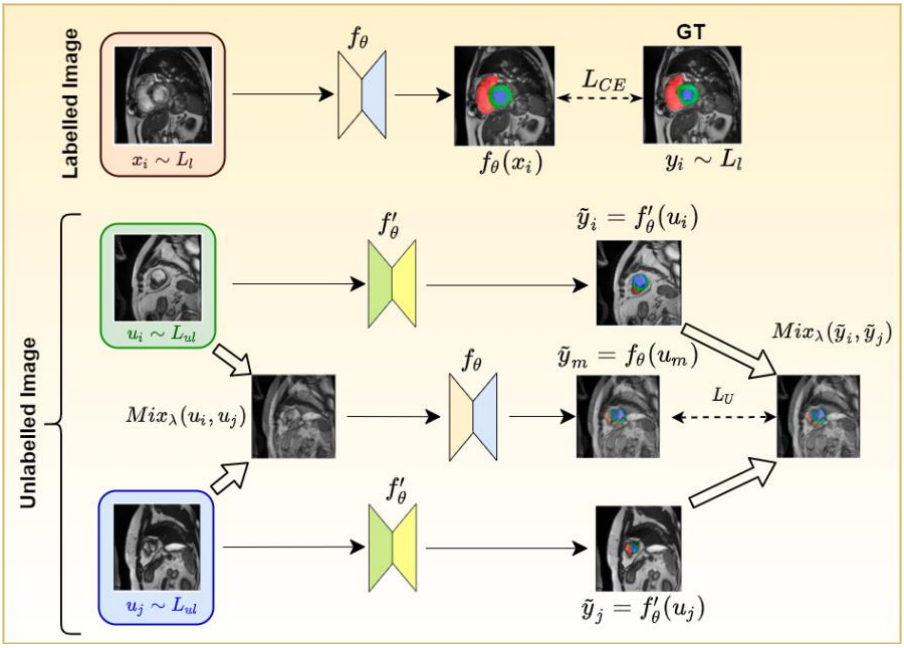

Fig. 1. Overall framework of the proposed architecture.

Thus, we have proposed a pixel-wise data perturbation strategy as consistency regularization in our work, which operates as described. The network is trained in such a way so as to ensure stable and accurate segmentation of image points interpolated from existing points. Considering two unlabeled image data-points u1 and u2, we interpolate another unlabeled image point Mα(u1; u2), where $M\alpha(u1; u2) = \alpha u1 + (1 - \alpha)u2$, for some hyperparameter α. Now, consistency regularization is applied between the output of the interpolated image data-point $f(M\alpha(u1; u2))$ and the interpolation of the outputs of the original unlabeled points $M\alpha(f(u1); f(u2)) = \alpha f(u1) + (1 - \alpha)f(u2)$. This exploits the fact that the network learns to predict a pixel-level segmentation mask of the input images, and further, consistency is maintained between the outputs of the interpolated inputs and the interpolated outputs of the original inputs. Thus, the unlabeled samples in the datasets are used to generate the new interpolated images and the corresponding pseudo-labels. It takes two unlabeled images as input and returns the interpolated image and the corresponding pseudo label, which is used by the network pipeline. Therefore, the Consistency Regularization technique can be summarized as:

$$M\alpha(f_{\theta'}(u1), f_{\theta'}(u2)) \cong f_{\theta}(M_{\alpha}(u1,u2)) \qquad (1)$$

This data mixing technique would help the model learn more robust features improving the semi-supervised learning on subsequent (target) tasks since random perturbations do not guarantee adversarial perturbation.

2.3 Objective Function

Consider labelled samples $(x_i,y_i) \sim L_l$ from joint distribution $P(X,Y)$ and unlabeled samples $(u_i, u_j) \sim L_{ul}$ from borderline distribution $P(X) = P(X,Y)/P(X|Y)$. Using SGD for every iteration t, the encoder-decoder parameter θ is updated
minimizing the objective function:

$$L = L_{CE} + r(t).L_U \qquad (2)$$

where L_{CE} is the cross-entropy loss applied over the labelled
data L_l and L_U is the interpolation consistency regularization loss applied over the unlabeled data L_{ul}, $r(t)$ is the ramp function adjusting the weight of L_U after every iteration. L_U is calculated over (u_i,u_j) of sampled mini batches and the pseudo labels $y_i = F_{\theta'}(u_i)$ and $y_j = F_{\theta'}(u_j)$ (θ' is the exponential moving average of θ). Next, interpolation $u_m = M\alpha(ui,uj)$ and model prediction $y_m = F_{\theta}(u_m)$ are computed updating θ to bring y_m closer to the interpolation of the pseudo labels, $M\alpha(y_i,y_j)$. The deviation in y_m and $M\alpha(y_i,y_j)$ is penalized using the mean squared loss. Therefore, L_U can be expressed as:

$$L_U = E_{ui,uj \sim L_{ul}} l(F_{\theta}(M\alpha(ui; uj)), M\alpha(F_{\theta'}(ui),F_{\theta'}(uj))) \qquad (3)$$

The overall approach is depicted in Figure 1.

3 Experiments and Results

3.1 Dataset and Experimental Setup

The method is originally applied on two public datasets: the ACDC 2017[6]and MMWHS dataset[7] In this paper, I applied this method to FLARE22 dataset after some modifications.

As mentioned above, The FLARE dataset is composed of a small number of labeled case(50) and a large number of unlabeled cases(2000) in the training set, 50 visible cases for validation, and 200 hidden cases for testing. The segmentation targets include 13 organs: liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum. But only 1000 unlabeled cases are used for semi-supervised training. And within labeled data, 10 case are split for validation and the left for labeled training.

We have used ResNet-50 as the encoder backbone of the architecture, using an ADAM optimizer with an initial learning rate of 1e-5. For evaluation purposes, three widely used metrics are used: Dice Similarity Score (DSC), Average Symmetric Distance(ASD), and Hausdorff Distance (HD). Average of all the metric scores over all the classes and reported in this paper. Mixing parameter α was set experimentally.

Algorithm 1: Pseudo-code of our proposed method.

Input:
L_l: Distribution of labelled samples; L_{ul}: Distribution of unlabelled samples
Define:
$f_\theta(\cdot)$: Segmentation network with trainable parameter θ
$f_{\theta'(\cdot)}$: Segmentation network with parameter θ'- exponential moving average of θ
T: Total number of iterations; $r(t)$: ramp function
λ: exponential moving average change rate
$\mathcal{M}_\alpha(u_1, u_2) = \alpha u_1 + (1-\alpha)u_2$

while (t $\leq$ T) **do**
  Sample labelled mini-batch $\triangleright$ $\{(x_p, y_p)\}_{p=1}^{P} \sim L_l$
  Supervised CE loss $\triangleright$ $\mathcal{L}_{\mathcal{CE}}(\{(f_\theta(x_p), y_p)\}_{p=1}^{P})$
  Sample two unlabelled batches $\triangleright$ $\{(u_i, u_j)\}_{k=1}^{U} \sim L_{ul}$
  Generate pseudo-labels $\triangleright$ $\{\tilde{y_i}, \tilde{y_j}\}_{k=1}^{U} = \{f_{\theta'}(u_i, u_j)\}_{k=1}^{U}$
  Interpolation $\triangleright$ $u_m = \mathcal{M}_\alpha(u_i, u_j), y_m = \mathcal{M}_\alpha(\tilde{y_i}, \tilde{y_j})$
  Interpolated pseudo-label $\triangleright$ $\tilde{y}_m = f_\theta(u_m)$
  Unsupervised loss $\triangleright$ $\mathcal{L}_\mathcal{U} = MSE(\{y_m, \tilde{y}_m\}_{m=1}^{U})$
  Overall model loss $\triangleright$ $\mathcal{L} = \mathcal{L}_{\mathcal{CE}} + r(t)\mathcal{L}_\mathcal{U}$
  Gradient computation $\triangleright$ $\mathcal{G}_\theta \longleftarrow \mathcal{L} \cdot \nabla_\theta$
  Update parameter $\triangleright$ $\theta' \longleftarrow (1-\lambda)\theta + \lambda\theta'$
  $\theta \longleftarrow step(\mathcal{G}_\theta, \theta)$
end while
return θ

Algorithm 1 shows the pseudo-code of the method during training. Curiously, unlike other SSL procedure, this method does not pretrain backbone network with labeled data. Instead, they used a two-way batch sampler to extract batch from labeled and unlabeled data at the mean time with the same amount(12 slices for each in this case). Then as the Algorithm 1 shows, both supervised loss(CE loss) and unsupervised loss(MSE loss) is calculated and combined to make the overall loss. That to say, they trained the supervised and unsupervised part at the mean time. Therefore, the given iteration of training is counted on the labeled part. And for the unlabeled part and the consistency regularization just iterated as much as needed. However, according to the poor result on FLARE22 dataset, whether this training procedure can improve efficiency of training without losing too much accuracy needs to be studied on other dataset. At least, it seems to work well on its original dataset(ACDC and MMWHS)

3.2 Results on the FLARE22 dataset

The validation results shows that it can yield 0.4120 DSC accuracy on liver segmentation, but on other organs, the DSC score are all under 0.10 except for 0.1070 on aorta. Whether the validation result is accurate needs to be further investigated. But based on the current validation, it suggests that this 2d consistency regularization with interpolation between two unlabeled slices is unsuitable for FLARE22 dataset or its task. However, it does appear to be efficient and have low consumption(RAM<2G, GPU<2G) as expected. Here is an example of the segmentation result of validation cases.

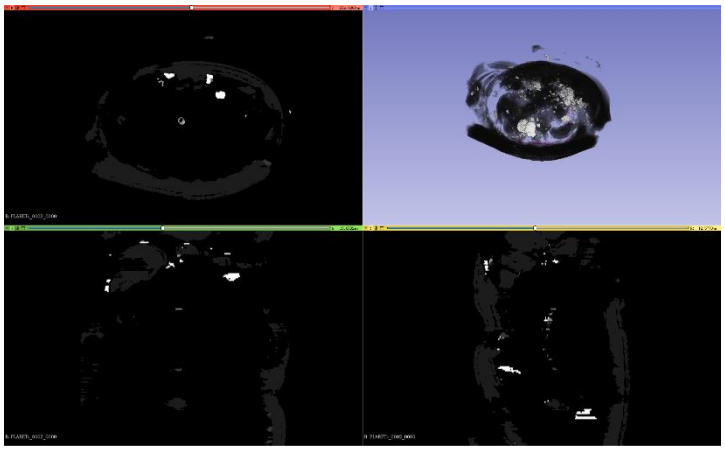

Fig. 2. An example of segmentation result

3.2 Discussion

In the original paper, this SSL method achieves DSC of 73.56%, 79.05%, and 89.80% on the ACDC 2017 dataset for 1.25%, 2.5%, and 10% labelled volumes respectively.

There are some possible reasons to account for the poor performance on the FLARE22 dataset. First of all, This method is a pure 2d method so that cannot capture inter-slice information of the 3d file. However, the dataset it originally applied on is also 3d medical images. The reasons maybe that the segmentation categories are less in ACDC dataset(3 targets plus background). And in FLARE22 dataset, segmentation categories expand into 13 organs including some minor ones. Thus, this simple interpolation and consistency regularization method can't handle the significant increase of difficulty. Besides, the original dataset is consisted of MRI data, so the z-dimension is relatively smaller and more invariant for segmentation targets. And since the target organs in FLARE22 dataset are more variant in z-dimension and the unlabeled slices are chose randomly for interpolation, it significantly weakens the effect of consistency regularization.

Based on the current results, I believe that to adapt for the FLARE22 dataset, introduction of more inter-slice information can improve the result significantly. As all the samples fed into the model is resampled into same sizes, executing a positional encoding to guide the interpolation of two unlabeled slices can be of help. With the help of positional encoding, we can make sure the interpolation is carried out between two slices in the same phase of each case. Further experiments need to be done to confirm whether this simple SSL method can be used for 3d CT dataset with complicated segmentation targets like FLARE22.

4 Conclusion

Here I have tried to apply a semi-supervised learning strategy that encourages consistency regularization by interpolation for segmentation of FLARE22. The original paper suggests that is advantageous over the previous SSL models in multiple aspects: unlike adversarial perturbations or generative models, it requires almost no additional computations. When applied to FLARE22 dataset, it does show high-efficiency and have low consumption. However, it fails to meet the segmentation accuracy standard. The main reason is its over-simplified 2d consistency regularization is unable to adjust complicated 3d CT dataset and multiple organs. More experiments need to be conducted to confirm whether this method can be used for complex tasks like FLARE22 after major modifications.

References

1. Bortsova, G. et al. (2019) 'Semi-supervised Medical Image Segmentation via Learning Consistency Under Transformations', in Medical Image Computing and Computer As-

sisted Intervention – MICCAI 2019. [Online]. Cham: Springer International Publishing. pp. 810–818.

2. X. Li, L. Yu, H. Chen, C. Fu, L. Xing, and P. Heng,"Transformation-consistent self-ensembling model for semi-supervised medical image segmentation," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 2, pp. 523–534,2020.
3. D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow, "Understanding and improving interpolation in autoencoders via an adversarial regular izer," arXiv preprint arXiv:1807.07543, 2018.
4. Basak, H. et al. (2022) An Embarrassingly Simple Consistency Regularization Method for Semi-Supervised Medical Image Segmentation.
5. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2020). nnU-Net: a self-configuring method or deep learning-based biomedical image segmentation. Nature Methods, 1-9.
6. Bernard, O. et al. (2018) Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE transactions on medical imaging. [Online] 37 (11), 2514–2525.
7. Zhuang, X. & Shen, J. (2016) Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Medical image analysis. [Online] 3177–87.