Medical Image Segmentation with Stochastic Aggregated Loss in a Unified U-Net

Phi Xuan Nguyen[1], Zhongkang Lu[2], Weimin Huang[2*], Su Huang[2], Akie Katsuki[3], Zhiping Lin[1]
[1]*School of EEE, Nanyang Technological University, Singapore*
[2]*Institute for Infocomm Research, Singapore*
[3]*GE Healthcare Japan Corporation, Japan*
[*]*Corresponding author: wmhuang@i2r.a-star.edu.sg*

***Abstract*—Automatic segmentation of medical images, such as computed tomography (CT) or magnetic resonance imaging (MRI), plays an essential role in efficient clinical diagnosis. While deep learning have gained popularity in academia and industry, more works have to be done to improve the performance for clinical practice. U-Net architecture, along with Dice coefficient optimization, has shown its effectiveness in medical image segmentation. Although it is an efficient measurement of the difference between the ground truth and the network's output, the Dice loss struggles to train with samples that do not contain targeted objects. While the situation is unusual in standard datasets, it is commonly seen in clinical data, where many training data available without the anomalies shown in the images, such as lesions and anatomic structures in some CTs/regions. In this paper, we propose a novel loss function - *Stochastic Aggregated Dice Coefficient (SA Dice)* and a modification of the network structure to improve its performance. Experimentally, in our own heart aorta CT dataset, our models beats the baseline by 4% in cross-validation Dice scores. In BRATS 2017 brain tumor segmentation challenge, the models also perform better than the state-of-the-art by approximately 2%.**

***Index Terms*—U-Net, CT, MRI, Dice, Segmentation, Stochastic Aggregated Loss**

I. Introduction

Imaging technologies, such as computed tomography (CT) or magnetic resonance imaging (MRI), are non-invasive methods for screening and diagnosis that have transformed the medical industry. As a result, automatic medical image segmentation become an important research topic. Fast and accurate segmentation helps physicians make more efficient and effective decisions in diagnosis and treatment plan. Therefore, there is an increasing demand for machine learning based tools capable of performing precise segmentation of medical images. However, problems specific to healthcare have limited the development of such automation technologies. For example, collecting and curating data is difficult and expensive as most of clinical data are not annotated, or only a small number of data are annotated. Manual labeling is tedious, exhaustive and requires professional skills to perform, especially for segmentation task. For diseased cases, the data can be scarce due to the limited number of patients available in an individual hospital.

Published in the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)

Recent studies on convolutional neural networks and deep learning [1] have shown tremendous success in medical image segmentation [2]–[5]. U-Net architecture, first proposed in [2], has demonstrated outstanding effectiveness in medical images. Since then, multiple works have been done to solve various segmentation tasks based on U-Net. In addition, Dice coefficient [6] has also become a standard for many optimization strategies for medical image segmentation [3]–[5], instead of pixel-wise cross entropy. Dice-based loss functions are particularly robust to class imbalance and shown very successful [3], [5], but they find examples without targets challenging because of gradient flatness problem that we will discuss later in section III-B.

This paper, based on U-Net models [3], [5], presents a solution to the aforementioned problem by proposing a novel yet simple loss function - *Stochastic Aggregated Dice Coefficient (SA Dice)* (section III-C). It also introduces *Weighted Multi-resolution Loss Component Accumulation (WMLA)* (section III-D) to achieve better performance. We tested the approaches in 2 datasets: Our own Cardiac Aorta CT images and BRATS 2017 Brain Tumor Segmentation Challenge. Experimentally, our methods surpass the baselines [3], [5], [7] by 4% in aorta segmentation and 2% in BRATS 2017 segmentation challenge [8].

II. Related Works

U-Net architecture was first introduced in [2] to tackle the microscopic ISBI challenge and it has gained tremendous success and popularity. Compared to other models proposed for natural image segmentation, such as MaskRCNN [9]; U-Net processes the image at different resolutions and feature channels as well as employs contraction-expansion residual connections. As a result, U-Net models are capable of recognizing complex micro features of medical images. Many variants of U-Net were proposed to address other segmentation problems. For example, V-Net [4] extends U-Net to three dimensions to solve MRI prostate cancer segmentation. AnatomyNet [7] also demonstrates effectiveness in MICCAI head and neck segmentation challenge 2015. In addition, the works in [3], [5] propose the use of residual connections [10] and element-wise summation of multiple segmentation maps from different resolutions. Their approaches have shown gains in hand and brain MRI image segmentation and BRATS 2017

challenge [8]. As such, our models are based on the U-Net networks proposed in [3], [5].

In terms of loss function, pixel-wise cross entropy was first used with U-Net in [2]. It was also used in M-Net [11] to solve MRI brain structure segmentation. However, since Dice score [6] becomes a regular metric fot this task, there has been many attempts to directly optimize this score by using Dice-based loss functions. For instance, V-Net [4] uses a novel Dice loss function while AnatomyNet [7] adopts a combination of Dice and focal loss. The works in [3], [5] also propose the use of Dice coefficient as the training objective function. In this paper, we investigate a potential limitation of many Dice-based loss functions proposed in such works (section III-B) and introduce a simple solution called Stochastic Aggregated Dice Coefficient (section III-C) to incorporate into our unified U-Net models.

III. Methods

A. Model Architecture

U-Net architecture comprises of two stages: contraction and expansion. It is also divided into N convolutional levels. In each level, there are multiple padded convolutional layers, a down-sampler (using max-pooling or strided convolution) and an up-sampler (using transpose convolution). During each level of contraction stage, U-Net models usually halve the image resolutions while double feature map channels. In expansion stage, conversely, the models restore the original resolutions and reduce the number of channels. In this paper, we adopt and modify the U-Net model proposed by [3] as it computes multi-level segmentation maps from the first three levels.

Not as [3], which use Jaccard index and for multiclass extension, we seek to maximize, through gradient ascent, the dice coefficient D between the predicted binary volume P and ground truth volume G. Since the optimization is stochastic, the overall loss is averaged across all examples in the batch of size M. Formally, let $p_{j,i} \in P_j$ and $g_{j,i} \in G_j$ be the predicted and ground truth voxels of the N-voxels image j in a batch of size M respectively, where $\forall p, g; p, g \in [0, 1]$. The stochastic dice coefficient D is defined as follow:

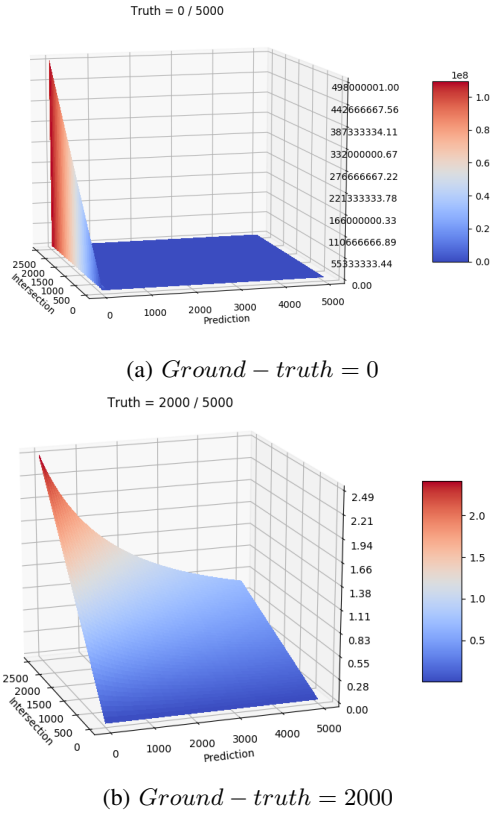
$$D = \frac{1}{M} \sum_{j}^{M} \frac{2 \sum_{i}^{N} p_{j,i} g_{j,i}}{\sum_{i}^{N} p_{j,i} + \sum_{i}^{N} g_{j,i}} \tag{1}$$

$$D_{multiclass} = \frac{1}{M} \sum_{j}^{M} \frac{1}{|K|} \sum_{k \in K} \frac{2 \sum_{i}^{N} p_{j,i,k} g_{j,i,k}}{\sum_{i}^{N} p_{j,i,k} + \sum_{i}^{N} g_{j,i,k}} \tag{2}$$

For the case of multiclass segmentation (more than one labels apart from the background), the dice coefficient is defined in equation (2); where $k \in K$ being the classes.

B. Gradient Flatness of Dice Coefficient in Extreme Conditions

Neural networks learn by optimizing an objective function through stochastic gradient descent/ascent algorithms. This requires the gradients of such loss function to be *smooth, continuous and non-zero*. In other words, the values of the loss must be *smooth, continuous and non-constant*. However, these characteristics of Dice loss in equation (1) vanishes when it faces extreme conditions of no-target labels. This is formally specified in equation (3) ; with p_i, g_i, N, D defined in section III-A.



(a) $Ground - truth = 0$

(b) $Ground - truth = 2000$

Fig. 1: The contour of dice coefficient (vertical axis) when ground-truth is zero (Fig. 1a) and 2000 / 5000 (Fig. 1b). The 2000 contour is smooth and continuous while the zero contour is sharp, flat and discontinuous.

$$\forall p_i | \sum_{i}^{N} p_i \neq 0, \lim_{g_i \to 0} D = \lim_{g_i \to 0} \frac{2 \sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i + \sum_{i}^{N} g_i} = 0 \tag{3}$$

Figure 1 also illustrates the problem. The contour of dice coefficient maintains a smooth continuity when the target-label amount is abundant, but becomes sharp and flat when the target-label approach zero. Zero loss produces no gradients to the model to update its parameters and causes the model trapped in local optima, thus makes it difficult to learn the data. We also found that the use of smooth Dice loss does not help relieve the issue. The proposed stochastic aggregated dice coefficient is designed to ease the impact of the problem.

C. Stochastic Aggregated Dice Coefficient

Instead of computing per-image dice coefficients and averaging them for each batch, we propose a new loss function: **Stochastic Aggregated Dice Loss (SA Dice)**, which differs from the original method. In this approach, all segmentation

outputs in a batch are merged into a large image and the dice loss is computed based on the aggregated image. Formally, let M, P, G respectively be the size of batch, the predicted binary volume and the ground truth volume. Let $p_{j,i} \in P_j$ and $g_{j,i} \in G_j$ be the predicted and ground truth voxels of image j in batch M respectively, where $p, g \in [0, 1] \forall p, g$. The proposed loss function D_{SA} is defined in equation (4):

$$D_{SA} = \frac{2\sum_j^M \sum_i^N p_{j,i} g_{j,i}}{\sum_j^M \sum_i^N p_{j,i} + \sum_j^M \sum_i^N g_{j,i}} \tag{4}$$

$$D_{SA,multiclass} = \frac{1}{|K|} \sum_{k \in K} \frac{2\sum_j^M \sum_i^N p_{j,i,k} g_{j,i,k}}{\sum_j^M \sum_i^N p_{j,i,k} + \sum_j^M \sum_i^N g_{j,i,k}} \tag{5}$$

For the case of multiclass segmentation, the SA dice coefficient is defined in equation (5); where $k \in K$ being the classes.

The rationale behind this approach is that the no-target examples are most likely to be accommodated with other examples that have labels during training. This will cause the aggregated dice coefficient contour smoother and curved due to the labels from such images. Therefore, the gradient signals will be noticeable and the model can be easily optimized. We also find experimentally that gradient clipping [12] and sometimes voxel-wise cross entropy loss beneficial.

D. Weighted Multi-resolution Loss Component Accumulation (WMLA)

Existing U-net frameworks compute the training loss between high-resolution ground-truths and the segmentation output of top level layer. In this section, we introduce **Weighted Multi-resolution Loss Component Accumulation (WMLA)** method, which computes loss components at different resolutions. The per-resolution component is calculated from segmentation output of each level and its respective down-sampled ground-truths. For example, suppose the output at level 2 is $32 * 32 * 32$, the ground-truth is down-sampled to resolution $32 * 32 * 32$ and dice coefficient component is computed at the level 2. The total loss of the model is the weighted sum of such loss components. This approach is illustrated in figure 2.

Formally, let P_k, G_k and $\alpha_k$be the predicted and ground truth image and per-resolution loss coefficient at level (resolution) k. The $Loss(P, Q)$ be the selected dice coefficient loss function. This can be equations (1), (4) or other loss function. The total multi-resolution loss is defined as in equation (6):

$$D_{multi-res} = \sum_k \alpha_k Loss(P_k, G_k) \tag{6}$$

In our experiments, the coefficient α_k remains as a hyper-parameter or a trainable parameter, but α_1 is kept constant at unity. Empirically, we found it beneficial to choose diminishing coefficients for lower levels or to use trainable parameter. In particular, we set $\alpha = (1, 0.1, 0.01)$ for the first, second and third levels. Using trainable α_k also produces performance gain.

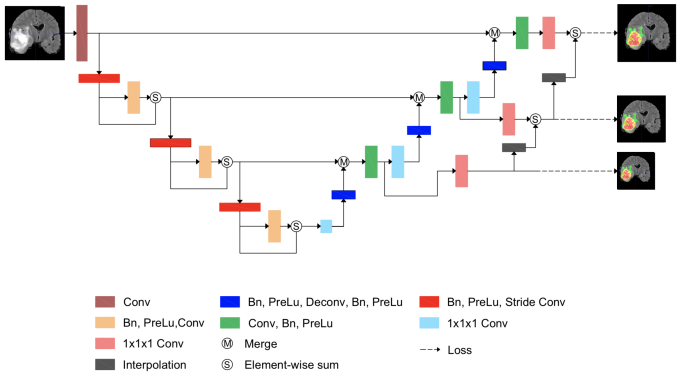


Fig. 2: The proposed U-Net architecture. It has additional loss components at multiple resolutions.

IV. Experiments

A. Setup

Similarly to [3], Our trained U-Net models, apart from the proposed changes, use 4 levels of contractions and expansions; 16 feature channels in the first level and the amount doubles when the resolutions are halved. We use Instance Normalization [13], Leaky ReLu activation and residual connections [10]. We used Adam optimizer [14] with learning rate $5e{-}4$ and trained for 300 epochs. In addition, we halved the learning rate when the validation loss no longer decreases. Due to memory limits, only batch size of 2 is used so that *SA Dice* approach can work properly. Finally, we trained our models with an Nvidia GTX 1080 Ti graphic card.

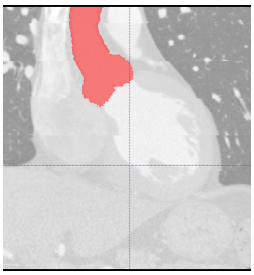
B. Datasets

Aorta is the largest artery in the human body, originated from the left ventricle. Successful aorta segmentation is essential for automatic and accurate discrimination of various vascular organs, such as the heart or coronary arteries. Our own aorta segmentation dataset contains 3D CT Coronary Angiography (CTCA) scans of the aorta from 50 patients. The 3D images are rescaled to $128 * 128 * 128$ resolution. They are split into training set and validation set by 90% and 10% ratios respectively. We performed several rounds of cross-validation and average the validation performance between all trials.

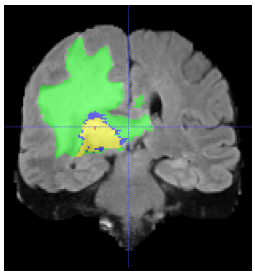
On the other hand, the Brain Tumor Segmentation dataset (BRATS 2017) [8] consists of 274 3D images with brain tumors. Similarly, such images are resized to $128 * 128 * 128$ resolution. They are also split into training set and validation set by 90% (246 samples) and 10% (28) ratios respectively. Similarly, cross validation is also performed on this dataset.

C. Results

1) Cardiac Aorta Segmentation: Table I shows experimental results for the aorta segmentation. As it can be seen, The U-Net model using SA Dice loss function achieves 0.9 dice score, exceeding the baseline score of 0.875 by 0.025 Dice points. It shows the effectiveness of the method. SA Dice improves the gradient signatures of examples with scarce target-label, reducing the amount of false-positives in such images. It also



(a) Cardiac Aorta



(b) BRATS Brain Tumor

Fig. 3: Cardiac aorta image and BRATS 2017 brain tumor sample.

TABLE I: Cross-validated results for **Cardiac Aorta** segmentation task.

Model	WMLA α	Dice
U-net (baseline) [3]	N.A	0.873
AnatomyNet [7]	N.A	0.874
SA Dice	N.A	0.900
SA Dice + WMLA	$\alpha = (1, 1, 1)$	0.908
SA Dice + WMLA	$\alpha = (1, 0.1, 0.01)$	**0.915**
SA Dice + WMLA	Trainable α	0.909

acts as an augmentation technique as it dynamically groups multiple combinations of training images. We also found that L2-norm gradient clipping of 5.0 or 3.0 beneficial in our experiments. In addition, the proposed *weighted multi-resolution loss component accumulation (WMLA)* also offers improvements compared to the baseline. This approach, when combined with SA Dice, achieves up to 0.915 dice score (0.04 points higher than the baseline). The results suggest that using diminishing α is crucial for the performance gain. This is justifiable because lower-level segmentation maps have low resolutions and are often less accurate and detailed. On the other hand, it focuses more on the overall loss at the object level.

2) BRATS 2017 Brain Tumor Segmentation: Experimental results for Brain Tumor segmentation task (BRATS 2017) [8] are shown in table II. Unlike the results for aorta dataset, only using SA Dice does not produce any noticeable performance gain while incorporating *weighted multi-resolution loss component accumulation (WMLA)* (section III-D) contributes to some improvements. To be more precise, a combination of multi-resolution losses with diminishing α factors, gradient clipping and voxel-wise cross entropy components reports a performance of 0.905 dice for whole tumor, 0.79 for core tumor and 0.754 for enhanced tumor segmentation benchmark. Compared to the baseline, this shows 0.015, 0.02 and 0.022 performance gains respectively.

TABLE II: Cross-validated results for **BRATS 2017 Brain Tumor** segmentation task.

Model	WMLA α	Dice		
		Whole	Core	Enhanced
U-net (baseline) [3]	N.A	0.890	0.770	0.732
AnatomyNet [7]	N.A	0.896	0.780	0.73
SA Dice	N.A	0.894	0.774	0.739
SA Dice + WMLA	$\alpha = (1, 1, 1)$	**0.907**	0.790	0.747
SA Dice + WMLA	$\alpha = (1, 0.1, 0.01)$	0.905	0.790	**0.754**
SA Dice + WMLA	Trainable α	0.902	**0.793**	0.750

V. Conclusions

In this paper, we investigated the limitation of existing Dice-based loss functions in extreme cases of medical image segmentation. We presented two methods to tackle the issue: Stochastic Aggregated Dice Coefficient (SA Dice) and Weighted Multi-resolution Loss Component Accumulation (WMLA). Our experiments show improvements of 4% and 2% for the cardiac aorta from CTCA and BRATS 2017 brain tumor segmentation from MRI compared to the baselines U-Net [3] and AnatomyNet [7].

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data," *CoRR*, vol. abs/1701.03056, 2017. [Online]. Available: http://arxiv.org/abs/1701.03056

[4] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[5] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge," *CoRR*, vol. abs/1802.10508, 2018. [Online]. Available: http://arxiv.org/abs/1802.10508

[6] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[7] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical physics*, 2018. [Online]. Available: http://arxiv.org/abs/1808.05238

[8] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, p. 1993, 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] R. Mehta and J. Sivaswamy, "M-net: A convolutional neural network for deep brain structure segmentation," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 437–440.

[12] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *CoRR*, vol. abs/1211.5063, 2012. [Online]. Available: http://arxiv.org/abs/1211.5063

[13] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016. [Online]. Available: http://arxiv.org/abs/1607.08022

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.