3D Segmentation with Exponential Logarithmic Loss for Highly Unbalanced Object Sizes*

Ken C. L. Wong, Mehdi Moradi**, Hui Tang, and Tanveer Syeda-Mahmood

IBM Research – Almaden Research Center, San Jose, CA, USA

Abstract. With the introduction of fully convolutional neural networks, deep learning has raised the benchmark for medical image segmentation on both speed and accuracy, and different networks have been proposed for 2D and 3D segmentation with promising results. Nevertheless, most networks only handle relatively small numbers of labels (<10), and there are very limited works on handling highly unbalanced object sizes especially in 3D segmentation. In this paper, we propose a network architecture and the corresponding loss function which improve segmentation of very small structures. By combining skip connections and deep supervision with respect to the computational feasibility of 3D segmentation, we propose a fast converging and computationally efficient network architecture for accurate segmentation. Furthermore, inspired by the concept of focal loss, we propose an exponential logarithmic loss which balances the labels not only by their relative sizes but also by their segmentation difficulties. We achieve an average Dice coefficient of 82% on brain segmentation with 20 labels, with the ratio of the smallest to largest object sizes as 0.14%. Less than 100 epochs are required to reach such accuracy, and segmenting a 128×128×128 volume only takes around 0.4 s.

1 Introduction

With the introduction of fully convolutional neural networks (CNNs), deep learning has raised the benchmark for medical image segmentation on both speed and accuracy [7]. Different 2D [5,8] and 3D [1,6,2,10] networks were proposed to segment various anatomies such as the heart, brain, liver, and prostate from medical images. Regardless of the promising results of these networks, 3D CNN image segmentation is still challenging. Most networks were applied on datasets with small numbers of labels (<10) especially in 3D segmentation. When more detailed segmentation is required with much more anatomical structures, previously unseen issues, such as computational feasibility and highly unbalanced object sizes, need to be addressed by new network architectures and algorithms.

* This paper was accepted by the International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018 (oral presentation). The final publication is available at Springer via https://doi.org/10.1007/978-3-030-00931-1_70

** Corresponding author.
There are only a few frameworks proposed for highly unbalanced labels. In [8], a 2D network architecture was proposed to segment all slices of a 3D brain volume. Error corrective boosting was introduced to compute label weights that emphasize parameter updates on classes with lower validation accuracy. Although the results were promising, the label weights were only applied to the weighted cross-entropy but not the Dice loss, and the stacking of 2D results for 3D segmentation may result in inconsistency among consecutive slices.

In [9], the generalized Dice loss was used as the loss function. Instead of computing the Dice loss of each label, the weighted sum of the products over the weighted sum of the sums between the ground-truth and predicted probabilities was computed for the generalized Dice loss, with the weights inversely proportional to the label frequencies. In fact, the Dice coefficient is unfavorable to small structures as a few pixels of misclassification can lead to a large decrease of the coefficient, and this sensitivity is irrelevant to the relative sizes among structures. Therefore, balancing by label frequencies is nonoptimal for Dice losses.

To address the issues of highly unbalanced object sizes and computational efficiency in 3D segmentation, we have two key contributions in this paper. I) We propose the exponential logarithmic loss function. In [4], to handle the highly unbalanced dataset of a two-class image classification problem, a modulating factor computed solely from the softmax probability of the network output is multiplied by the weighted cross-entropy to focus on the less accurate class. Inspired by this concept of balancing classification difficulties, we propose a loss function comprising the logarithmic Dice loss which intrinsically focuses more on less accurately segmented structures. The nonlinearities of the logarithmic Dice loss and the weighted cross-entropy can be further controlled by the proposed exponential parameters. In this manner, the network can achieve accurate segmentation on both small and large structures. II) We propose a fast converging and computationally efficient network architecture by combining the advantages of skip connections and deep supervision, which has only about 1/14 of the parameters of, and is twice as fast as, the V-Net [6]. Experiments were performed on brain magnetic resonance (MR) images with 20 highly unbalanced labels. Combining these two innovations achieved an average Dice coefficient of 82% with the average segmentation time as 0.4 s.

2 Methodology

2.1 Proposed Network Architecture

3D segmentation networks require much more computational resources than 2D networks. Therefore, we propose a network architecture which aims at accurate segmentation and fast convergence with respect to limited resources (Fig. 1). Similar to most segmentation networks, our network comprises the encoding and decoding paths. The network is composed of convolutional blocks, each comprises \( k \) cascading \( 3 \times 3 \times 3 \) convolutional layers of \( n \) channels associated with batch normalization (BN) and rectified linear units (ReLU). For better convergence, a skip connection with a \( 1 \times 1 \times 1 \) convolutional layer is used in each block.
Instead of concatenation, we add the two branches together for less memory consumption, so the block allows efficient multi-scale processing and deeper networks can be trained. The number of channels \( n \) is doubled after each max pooling and is halved after each upsampling. More layers \( k \) are used with tensors of smaller sizes so that more abstract knowledge can be learned with feasible memory use. Feature channels from the encoding path are concatenated with the corresponding tensors in the decoding path for better convergence. We also include a Gaussian noise layer and a dropout layer to avoid overfitting.

Similar to \[5\], we utilize deep supervision which allows more direct backpropagation to the hidden layers for faster convergence and better accuracy \[3\]. Although deep supervision significantly improves convergence, it is memory expensive especially in 3D networks. Therefore, we omit the tensor from the block with the most channels (Block(192, 3)) so that training can be performed on a GPU with 12 GB of memory. A final layer of 1\( \times \)1\( \times \)1 convolution with the softmax function provides the segmentation probabilities.

### 2.2 Exponential Logarithmic Loss

We propose a loss function which improves segmentation on small structures:

\[
L_{\text{Exp}} = w_{\text{Dice}}L_{\text{Dice}} + w_{\text{Cross}}L_{\text{Cross}}
\]

with \( w_{\text{Dice}} \) and \( w_{\text{Cross}} \) the respective weights of the exponential logarithmic Dice loss \( L_{\text{Dice}} \) and the weighted exponential cross-entropy \( L_{\text{Cross}} \):

\[
L_{\text{Dice}} = E\left[\left(-\ln(Dice_{i}(x))\right)^{\gamma_{\text{Dice}}}\right] \quad \text{with} \quad Dice_{i} = \frac{2\left(\sum_{x}\delta_{il}(x)p_{i}(x)\right) + \epsilon}{\left(\sum_{x}\delta_{il}(x) + p_{i}(x)\right) + \epsilon}
\]

\[
L_{\text{Cross}} = E\left[w_{l}\left(-\ln(p_{l}(x))\right)^{\gamma_{\text{Cross}}}\right]
\]

with \( x \) the pixel position and \( i \) the label. \( l \) is the ground-truth label at \( x \). \( E[\cdot] \) is the mean value with respect to \( i \) and \( x \) in \( L_{\text{Dice}} \) and \( L_{\text{Cross}} \), respectively. \( \delta_{il}(x) \) is the Kronecker delta which is 1 when \( i = l \) and 0 otherwise. \( p_{i}(x) \) is the softmax probability which acts as the portion of pixel \( x \) owned by label \( i \) when computing...
Fig. 2. Loss functions with different nonlinearities, where \( x \) can be Dice\(_i\) or \( p_l(x) \).

Dice\(_i\). \( \epsilon = 1 \) is the pseudocount for additive smoothing to handle missing labels in training samples. \( w_l = ((\sum_k f_k)/f_l)^{0.5} \), with \( f_k \) the frequency of label \( k \), is the label weight for reducing the influences of more frequently seen labels. \( \gamma_{\text{Dice}} \) and \( \gamma_{\text{Cross}} \) further control the nonlinearities of the loss functions, and we use \( \gamma_{\text{Dice}} = \gamma_{\text{Cross}} = \gamma \) here for simplicity.

The use of the Dice loss in CNN was proposed in [6]. The Dice coefficient is unfavorable to small structures as misclassifying a few pixels can lead to a large decrease of the coefficient. The use of label weights cannot alleviate such sensitivity as it is irrelevant to the relative object sizes, and the Dice coefficient is already a normalized metric. Therefore, instead of size differences, we use the logarithmic Dice loss which focuses more on less accurate labels. Fig. 2 shows a comparison between the linear (\( E[1 - \text{Dice}_i] \)) and logarithmic Dice loss.

We provide further control on the nonlinearities of the losses by introducing the exponents \( \gamma_{\text{Dice}} \) and \( \gamma_{\text{Cross}} \). In [4], a modulating factor, \((1 - p_l)^\gamma\), is multiplied by the weighted cross-entropy to become \( w_l(1 - p_l)^\gamma(-\ln(p_l)) \) for two-class image classification. Apart from balancing the label frequencies using the label weights \( w_l \), this focal loss also balances between easy and hard samples. Our exponential loss achieves a similar goal. With \( \gamma > 1 \), the loss focuses more on less accurate labels than the logarithmic loss (Fig. 2). Although the focal loss works well for the two-class image classification in [4], we got worse results when applying to our segmentation problem with 20 labels. This may be caused by the over suppression of the loss function when the label accuracy becomes high. In contrast, we could get better results with \( 0 < \gamma < 1 \). Fig. 2 shows that when \( \gamma = 0.3 \), there is an inflection point around \( x = 0.5 \), where \( x \) can be Dice\(_i\) or \( p_l(x) \). For \( x < 0.5 \), this loss behaves similarly to the losses with \( \gamma \geq 1 \) with decreasing gradient magnitude as \( x \) increases. This trend reverses for \( x > 0.5 \) with increasing gradient magnitude. In consequence, this loss encourages improvements at both low and high prediction accuracy. This characteristics is the reason of using the proposed exponential form instead of the one in [4].

2.3 Training Strategy

Image augmentation is used to learn invariant features and avoid overfitting. As realistic nonrigid deformation is difficult to implement and computationally
expensive, we limit the augmentation to rigid transformations including rotation (axial, $\pm 30^\circ$), shifting ($\pm 20\%$), and scaling ($[0.8, 1.2]$). Each image has an $80\%$ chance to be transformed in training, thus the number of augmented images is proportional to the number of epochs. The optimizer Adam is used with the Nesterov momentum for fast convergence, with the learning rate as $10^{-3}$, batch size as one, and 100 epochs. A TITAN X GPU with 12 GB of memory is used.

3 Experiments

3.1 Data and Experimental Setups

A dataset of 43 3D brain MR images from different patients was neuroanatomically labeled to provide the training and validation samples. The images were produced by the T1-weighted MP-RAGE pulse sequence which provides high tissue contrast. They were manually segmented by highly trained experts with the results reviewed by a consulting neuroanatomist. Each segmentation had 19 semantic labels of brain structures, thus 20 labels with the background included (Table 1(a)). As there were various image sizes (128 to 337) and spacings (0.9 to 1.5 mm), each image was resampled to isotropic spacing using the minimum spacing, zero padded on the shorter sides, and resized to $128 \times 128 \times 128$.

Table 1(a) shows that the labels were highly unbalanced. The background occupied 93.5\% of an image on average. Without the background, the relative sizes of the smallest and largest structures were 0.07\% and 50.24\%, respectively, thus a ratio of 0.14\%.

We studied six loss functions using the proposed network, and applied the best one to the V-Net architecture [6], thus a total of seven cases were studied. For $L_{Exp}$, we set $w_{\text{Dice}} = 0.8$ and $w_{\text{Cross}} = 0.2$ as they provided the best results. Five sets of data were generated by shuffling and splitting the dataset, with 70\% for training and 30\% for validation in each set. Experiments were performed on all five sets of data for each case studied for more statistically sound results. The actual Dice coefficients, not the Dice $i$ in (2), were computed for each validation image. Identical setup and training strategy were used in all experiments.

3.2 Results and Discussion

Table 1(b) shows the Dice coefficients averaged from the five experiments. The linear Dice loss ($E[1 - \text{Dice}_i]$) had the worst performance. It performed well with the relatively large structures such as the gray and white matters, but the performance decreased with the sizes of the structures. The very small structures, such as the nucleus accumbens and amygdala, were missed in all experiments. In contrast, the logarithmic Dice loss ($L_{\text{Dice}}(\gamma = 1)$) provided much better results, though the large standard deviation of label 2 indicates that there were misses. We also performed experiments with the weighted cross-entropy ($L_{\text{Cross}}(\gamma = 1)$), whose performance was better than the linear Dice loss but worse than the logarithmic Dice loss. The weighted sum of the logarithmic Dice loss and
Table 1. Semantic brain segmentation. (a) Semantic labels and their relative sizes on average (%) without the background. CVL represents cerebellar vermal lobules. The background occupied 93.5% of an image on average. (b) Dice coefficients between prediction and ground truth averaged from five experiments (format: mean±std%). The best results are highlighted in blue. $w_{\text{Dice}} = 0.8$ and $w_{\text{Cross}} = 0.2$ for all $L_{\text{Exp}}$.

(a) Semantic labels and their relative sizes on average (%).

| Label | Percentage |
|-------|------------|
| 1. Cerebral grey | 50.24 |
| 2. 3rd ventricle | 0.09 |
| 3. 4th ventricle | 0.15 |
| 4. Brainstem | 1.46 |
| 5. CVL I-V | 0.39 |
| 6. CVL VI-VII | 0.19 |
| 7. CVL VIII-X | 0.26 |
| 8. Accumbens | 0.07 |
| 9. Amygdala | 0.21 |
| 10. Caudate | 0.54 |
| 11. Cerebellar grey | 0.19 |
| 12. Cerebellar white | 0.06 |
| 13. Cerebral white | 31.23 |
| 14. Hippocampus | 0.58 |
| 15. Inf. lateral vent. | 0.09 |
| 16. Lateral ventricle | 2.11 |
| 17. Pallidum | 0.25 |
| 18. Putamen | 0.73 |
| 19. Thalamus | 1.19 |

(b) Average Dice coefficients (mean±std%) with respective to the ground truth.

| $E[1 - \text{Dice}]$ | 1.87 ± 9.75 |
| $L_{\text{Dice}}(\gamma = 1)$ | 1.87 ± 9.75 |
| $L_{\text{Cross}}(\gamma = 1)$ | 1.87 ± 9.75 |
| $L_{\text{Exp}}(\gamma = 1)$ | 1.87 ± 9.75 |
| $L_{\text{Exp}}(\gamma = 2)$ | 1.87 ± 9.75 |
| $L_{\text{Exp}}(\gamma = 0.3)$ | 1.87 ± 9.75 |

Proposed network with weighted cross-entropy ($L_{\text{Exp}}(\gamma = 1)$) outperformed the individual losses, and it provided the second best results among the tested cases. As discussed in Section 2.2, $L_{\text{Exp}}(\gamma = 2)$ was ineffective even on larger structures. This is consistent with our observation in Fig. 2 that the loss function is over suppressed when the accuracy is getting higher. In contrast, $L_{\text{Exp}}(\gamma = 0.3)$ gave the best results. Although it only performed slightly better than $L_{\text{Exp}}(\gamma = 1)$ in terms of the means, the smaller standard deviations indicate that it was also more precise.

When applying the best loss function to the V-Net, its performance was only better than the linear Dice loss and $L_{\text{Exp}}(\gamma = 2)$. This shows that our proposed network architecture performed better than the V-Net on this problem.

Fig. 8 shows the validation Dice coefficients vs. epoch, averaged from the five experiments. Instead of the losses, we show the Dice coefficients as their magnitudes were consistent among cases. Similar to Table 1(b), the logarithmic Dice
loss, $L_{\text{Exp}}(\gamma = 1)$, and $L_{\text{Exp}}(\gamma = 0.3)$ had good convergence and performance, with $L_{\text{Exp}}(\gamma = 0.3)$ performed slightly better. These three cases converged at about 80 epochs. The weighted cross-entropy and $L_{\text{Exp}}(\gamma = 2)$ were more fluctuating. The linear Dice loss also converged at about 80 epochs but with a much smaller Dice coefficient. Comparing between the V-Net and the proposed network with $L_{\text{Exp}}(\gamma = 0.3)$, the V-Net had worse convergence especially at the earlier epochs. This shows that the proposed network had better convergence.

Fig. 4 shows the visualization of an example. There are two obvious observations. First of all, consistent with Table 1(b), the linear Dice loss missed some small structures such as the nucleus accumbens and amygdala, though it performed well on large structures. Secondly, the segmentation of the V-Net deviated a lot from the ground truth. The logarithmic Dice loss, $L_{\text{Exp}}(\gamma = 1)$, and $L_{\text{Exp}}(\gamma = 0.3)$ had the best segmentations and average Dice coefficients. The weighted cross-entropy had the same average Dice coefficient as $L_{\text{Exp}}(\gamma = 2)$, though the weighted cross-entropy over-segmented some structures such as the brainstem, and $L_{\text{Exp}}(\gamma = 2)$ had a noisier segmentation.

Comparing the efficiencies between the proposed network and the V-Net, the proposed network had around 5 million parameters while the V-Net had around
71 million parameters, a 14-fold difference. Furthermore, the proposed network only took about 0.4 s on average to segment a 128×128×128 volume, while the V-Net took about 0.9 s. Therefore, the proposed network was more efficient.

4 Conclusion

In this paper, we propose a network architecture optimized for 3D image segmentation, and a loss function for segmenting very small structures. The proposed network architecture has only about 1/14 of the parameters of, and is twice as fast as, the V-Net. For the loss function, the logarithmic Dice loss outperforms the linear Dice loss, and the weighted sum of the logarithmic Dice loss and the weighted cross-entropy outperforms the individual losses. With the introduction of the exponential form, the nonlinearities of the loss functions can be further controlled to improve the accuracy and precision of segmentation.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS, vol. 9901, pp. 424–432 (2016)
2. Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.A.: 3D deeply supervised network for automated segmentation of volumetric medical images. Medical Image Analysis 41, 40–54 (2017)
3. Lee, C.Y., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: International Conference on Artificial Intelligence and Statistics. pp. 562–570 (2015)
4. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. arXiv:1708.02002 [cs.CV] (2017)
5. Mehta, R., Sivaswamy, J.: M-net: A convolutional neural network for deep brain structure segmentation. In: IEEE International Symposium on Biomedical Imaging. pp. 437–440 (2017)
6. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: IEEE International Conference on 3D Vision. pp. 565–571 (2016)
7. 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. LNCS, vol. 9351, pp. 234–241 (2015)
8. Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C.: Error corrective boosting for learning fully convolutional networks with limited data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS, vol. 10435, pp. 231–239 (2017)
9. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. LNCS, vol. 10553, pp. 240–248 (2017)
10. Tang, H., Moradi, M., El Harouni, A., Wang, H., Veni, G., Prasanna, P., Syeda-Mahmood, T.: Segmentation of anatomical structures in cardiac CTA using multi-label V-Net. In: Medical Imaging 2018: Image Processing. vol. 10574, p. 1057407 (2018)