Tversky loss function for image segmentation using 3D fully convolutional deep networks

Seyed Sadegh Mohseni Salehi[1,2⋆], Deniz Erdogmus[1], and Ali Gholipour[2]

[1] Electrical and Computer Engineering Department, Northeastern University
[2] Radiology Department, Boston Children's Hospital; and Harvard Medical School, Boston, MA, 02115,

Abstract. Fully convolutional deep neural networks carry out excellent potential for fast and accurate image segmentation. One of the main challenges in training these networks is data imbalance, which is particularly problematic in medical imaging applications such as lesion segmentation where the number of lesion voxels is often much lower than the number of non-lesion voxels. Training with unbalanced data can lead to predictions that are severely biased towards high precision but low recall (sensitivity), which is undesired especially in medical applications where false negatives are much less tolerable than false positives. Several methods have been proposed to deal with this problem including balanced sampling, two step training, sample re-weighting, and similarity loss functions. In this paper, we propose a generalized loss function based on the Tversky index to address the issue of data imbalance and achieve much better trade-off between precision and recall in training 3D fully convolutional deep neural networks. Experimental results in multiple sclerosis lesion segmentation on magnetic resonance images show improved F_2 score, Dice coefficient, and the area under the precision-recall curve in test data. Based on these results we suggest Tversky loss function as a generalized framework to effectively train deep neural networks.

1 Introduction

Deep convolutional neural networks have attracted enormous attention in medical image segmentation as they have shown superior performance compared to conventional methods in several applications. This includes automatic segmentation of brain lesions [2,10], tumors [9,15,21], and neuroanatomy [14,22,3], using voxelwise network architectures [14,9,17], and more recently using 3D voxelwise networks [3,10], and fully convolutional networks (FCNs) [4,13,17]. Compared to voxelwise methods, FCNs are fast in testing and training, and use the entire samples to learn local and global image features. On the other hand, voxelwise networks may use a subset of samples to reduce data imbalance issues and increase efficiency.

Data imbalance is a common issue in medical image segmentation. For example in lesion detection the number of non-lesion voxels is typically > 500 times

⋆ Corresponding author: S.S.M.Salehi (email: ssalehi@ece.neu.edu).

larger than the number of diagnosed lesion voxels. Without balancing the labels the learning process may converge to local minima of a sub-optimal loss function, thus predictions may strongly bias towards non-lesion tissue. The outcome will be high-precision, low-recall segmentations. This is undesired especially in computer-aided diagnosis or clinical decision support systems where high sensitivity (recall) is a key characteristic of an automatic detection system.

A common approach to account for data imbalance, especially in voxelwise methods, is to extract equal training samples from each class [20]. The downsides of this approach are that it does not use all the information content of the images and may bias towards rare classes. Hierarchical training [5,21,20] and retraining [9] have been proposed as alternative strategies but they can be prone to overfitting and sensitive to the state of the initial classifiers [10]. Recent training methods for FCNs resorted to loss functions based on sample re-weighting [2,10,12,16,18], where lesion regions, for example, are given more importance than non-lesion regions during training. In the re-weighting approach, to balance the training samples between classes, the total cost is calculated by computing the weighted mean of each class. The weights are inversely proportional to the probability of each class appearance, i.e. higher appearance probabilities lead to lower weights. Although this approach works well for some relatively unbalanced data like brain extraction [17] and tumor detection [15], it becomes difficult to calibrate and does not perform well for highly unbalanced data such as lesion detection. To eliminate sample re-weighting, Milletari et. al. proposed a loss function based on the Dice similarity coefficient [13].

The Dice loss layer is a harmonic mean of precision and recall thus weighs false positives (FPs) and false negatives (FNs) equally. To achieve a better trade-off between precision and recall (FPs vs. FNs), we propose a loss layer based on the Tversky similarity index [19]. Tversky index is a generalization of the Dice similarity coefficient and the F_β scores. We show how adjusting the hyperparameters of this index allow placing emphasis on false negatives in training a network that generalizes and performs well in highly imbalanced data as it leads to high sensitivity, Dice, F_2 score, and the area under the precision-recall (PR) curve [1] in the test set. To this end, we adopt a 3D FCN, based on the U-net architecture, with a Tversky loss layer, and test it in the challenging multiple sclerosis lesion detection problem on multi-channel MRI [6,20]. The ability to train a network for higher sensitivity (recall) in the expense of acceptable decrease in precision is crucial in many medical image segmentation tasks such as lesion detection.

2 Method

2.1 Network architecture

We design and evaluate our 3D fully convolutional network [12,18] based on the U-net architecture [16]. To this end, we develop a 3D U-net based on AutoNet [17] and introduce a new loss layer based on the Tversky index. This U-net style architecture, which has been designed to work with very small number of training images, is shown in Figure 1. It consists of a contracting path (to the right) and an expanding path (to the left). To learn and use local information,

high-resolution 3D features in the contracting path are concatenated with upsampled versions of global low-resolution 3D features in the expanding path. Through this concatenation the network learns to use both high-resolution local features and low-resolution global features.

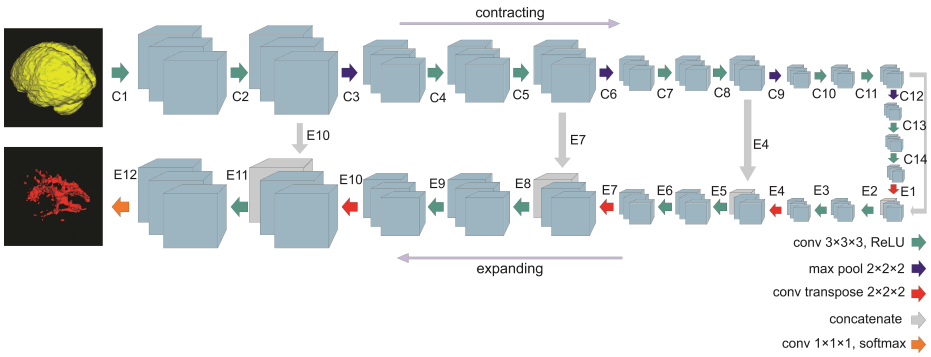

Fig. 1. The 3D U-net style architecture; The complete description of the input and output size of each level is presented in Table S1 in the supplementary material.

The contracting path contains padded $3\times3\times3$ convolutions followed by ReLU non-linear layers. A $2\times2\times2$ max pooling operation with stride 2 is applied after every two convolutional layers. After each downsampling by the max pooling layers, the number of features is doubled. In the expanding path, a $2 \times 2 \times 2$ transposed convolution operation is applied after every two convolutional layers, and the resulting feature map is concatenated to the corresponding feature map from the contracting path. At the final layer a $1\times1\times1$ convolution with softmax output is used to reach the feature map with a depth equal to the number of classes (lesion or non-lesion tissue), where the loss function is calculated. The size of the network layers is shown in Table S1 in the supplementary material.

2.2 Tversky loss layer

The output layer in the network consists of c planes, one per class ($c = 2$ in lesion detection). We applied softmax along each voxel to form the loss. Let P and G be the set of predicted and ground truth binary labels, respectively. The Dice similarity coefficient D between two binary volumes is defined as:

$$D(P,G) = \frac{2|PG|}{|P| + |G|} \tag{1}$$

If this is used in a loss layer in training [13], it weighs FPs and FNs (precision and recall) equally. In order to weigh FNs more than FPs in training our network for highly imbalanced data, where detecting small lesions is crucial, we propose a loss layer based on the Tversky index [19]. The Tiversky index is defined as:

$$S(P,G;\alpha,\beta) = \frac{|PG|}{|PG| + \alpha|P \setminus G| + \beta|G \setminus P|} \tag{2}$$

where α and β control the magnitude of penalties for FPs and FNs, respectively.

To define the Tversky loss function we use the following formulation:

$$T(\alpha,\beta)=\frac{\sum_{i=1}^{N}p_{0i}g_{0i}}{\sum_{i=1}^{N}p_{0i}g_{0i}+\alpha\sum_{i=1}^{N}p_{0i}g_{1i}+\beta\sum_{i=1}^{N}p_{1i}g_{0i}} \tag{3}$$

where in the output of the softmax layer, the p_{0i} is the probability of voxel i be a lesion and p_{1i} is the probability of voxel i be a non-lesion. Also, g_{0i} is 1 for a lesion voxel and 0 for a non-lesion voxel and vice verse for the g_{1i}. The gradient of the loss in Equation 3 with respect to p_{0i} and p_{1i} can be calculated as:

$$\frac{\partial T}{\partial p_{0i}}=2\frac{g_{0j}(\sum_{i=1}^{N}p_{0i}g_{0i}+\alpha\sum_{i=1}^{N}p_{0i}g_{1i}+\beta\sum_{i=1}^{N}p_{1i}g_{0i})-(g_{0j}+\alpha g_{1j})\sum_{i=1}^{N}p_{0i}g_{0i}}{(\sum_{i=1}^{N}p_{0i}g_{0i}+\alpha\sum_{i=1}^{N}p_{0i}g_{1i}+\beta\sum_{i=1}^{N}p_{1i}g_{0i})^2} \tag{4}$$

$$\frac{\partial T}{\partial p_{1i}}=-\frac{\beta g_{1j}\sum_{i=1}^{N}p_{0i}g_{0i}}{(\sum_{i=1}^{N}p_{0i}g_{0i}+\alpha\sum_{i=1}^{N}p_{0i}g_{1i}+\beta\sum_{i=1}^{N}p_{1i}g_{0i})^2} \tag{5}$$

Using this formulation we do not need to balance the weights for training. Also by adjusting the hyperparameters α and β we can control the trade-off between false positives and false negatives. It is noteworthy that in the case of $\alpha=\beta=0.5$ the Tversky index simplifies to be the same as the Dice coefficient, which is also equal to the F_1 score. With $\alpha=\beta=1$, Equation 2 produces Tanimoto coefficient, and setting $\alpha+\beta=1$ produces the set of F_β scores. Larger βs weigh recall higher than precision (by placing more emphasis on false negatives). We hypothesize that using higher βs in our generalized loss function in training will lead to higher generalization and improved performance for imbalanced data; and effectively helps us shift the emphasis to lower FNs and boost recall.

2.3 Experimental design

We tested our FCN with Tversky loss layer to segment multiple sclerosis (MS) lesions [6,20]. T1-weighted, T2-weighted, and FLAIR MRI of 15 subjects were used as input, where we used two-fold cross-validation for training and testing. Images of different sizes were all rigidly registered to a reference image at size $128\times224\times256$. Our 3D-Unet was trained end-to-end. Cost minimization on 1000 epochs was performed using ADAM optimizer [11] with an initial learning rate of 0.0001 multiplied by 0.9 every 1000 stjpg. The training time for this network was approximately 4 hours on a workstation with Nvidia Geforce GTX1080 GPU.

The test fold MRI volumes were segmented using feedforward through the network. The output of the last convolutional layer with softmax non-linearity consisted of a probability map for lesion and non-lesion tissues. Voxels with computed probabilities of 0.5 or more were considered to belong to the lesion tissue and those with probabilities <0.5 were considered non-lesion tissue.

2.4 Evaluation metrics

To evaluate the performance of the networks and compare them against state-of-the-art in MS lesion segmentation, we report Dice similarity coefficient (DSC):

$$DSC = \frac{2\,|P \cap R|}{|P| + |R|} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

where P and R are the predicted and ground truth labels, respectively; and TP, FP, and FN are the true positive, false positive, and false negative rates, respectively. We also calculate and report specificity, $\frac{TN}{TN+FP}$, and sensitivity, $\frac{TP}{TP+FN}$, and the F_2 score as a measure that is commonly used in applications where recall is more important than precision (as compared to F_1 or DSC):

$$F_2 = \frac{5TP}{5TP + 4FN + FP} \tag{7}$$

To critically evaluate the performance of the detection for the highly unbalanced (skewed) dataset, we use the Precision-Recall (PR) curve (as opposed to the receiver-operator characteristic, or ROC, curve) as well as the area under the PR curve (the APR score) [1,7,8]. For such skewed datasets, the PR curves and APR scores (on test data) are preferred figures of algorithm performance.

3 Results

To evaluate the effect of Tversky loss function and compare it with Dice in lesion segmentation, we trained our FCN with different α and β values. The performance metrics (on the test set) have been reported in Table 1. The results show that 1) the balance between sensitivity and specificity was controlled by the parameters of the loss function; and 2) according to all combined test measures, the best results were obtained from the FCN trained with $\beta = 0.7$, which performed much better than the FCN trained with the Dice loss layer corresponding to $\alpha = \beta = 0.5$.

Penalties	DSC	Sensitivity	Specificity	F_2 score	APR score
$\alpha = 0.5$, $\beta = 0.5$	53.42	49.85	**99.93**	51.77	52.57
$\alpha = 0.4$, $\beta = 0.6$	54.57	55.85	99.91	55.47	54.34
$\alpha = 0.3$, $\beta = 0.7$	**56.42**	56.85	99.93	**57.32**	**56.04**
$\alpha = 0.2$, $\beta = 0.8$	48.57	61.00	99.89	54.53	53.31
$\alpha = 0.1$, $\beta = 0.9$	46.42	**65.57**	99.87	56.11	51.65

Table 1. Performance metrics (on the test set) for different values of the hyperparameters α and β used in training the FCN. The best values for each metric have been highlighted in bold. As expected, it is observed that higher β led to higher sensitivity (recall) and lower specificity. The combined performance metrics, in particular APR, F_2 and DSC indicate that the best performance was achieved at $\beta = 0.7$. Note that the FCN trained with the Dice loss function ($\beta = 0.5$) did not generate good results.

Figure 2(a) shows the PR curve for the entire test dataset, and Figure 2(b) and (c) show the PR curves for two cases with extremely high, and extremely

low density of lesions, respectively. The best results based on the precision-recall trade-off were always obtained at $\beta = 0.7$ and not with the Dice loss function.

Figures 3 and 4 show the effect of different penalty magnitudes (βs) on segmenting a subject with high density of lesions, and a subject with very few lesions, respectively. These cases, that correspond to the PR curves shown in Figure 2(b and c), show that the best performance was achieved by using a loss function with $\beta = 0.7$ in training. We note that the network trained with the Dice loss layer ($\beta = 0.5$) did not detect the lesions in the case shown in Figure 4.

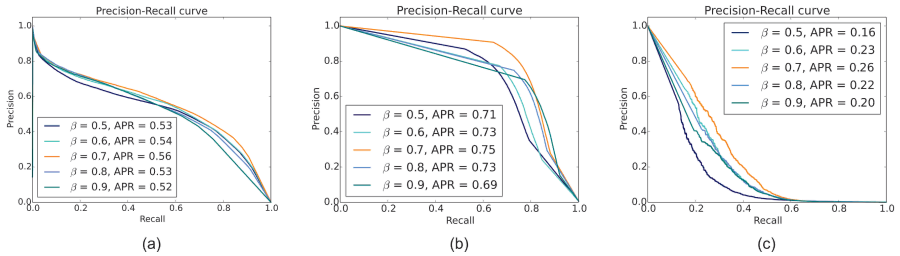

Fig. 2. PR curves with different α and β for: (a) all test set; (b) a subject with high density of lesions (Fig. 3); and (c) a subject with very low density of lesions (Fig. 4).

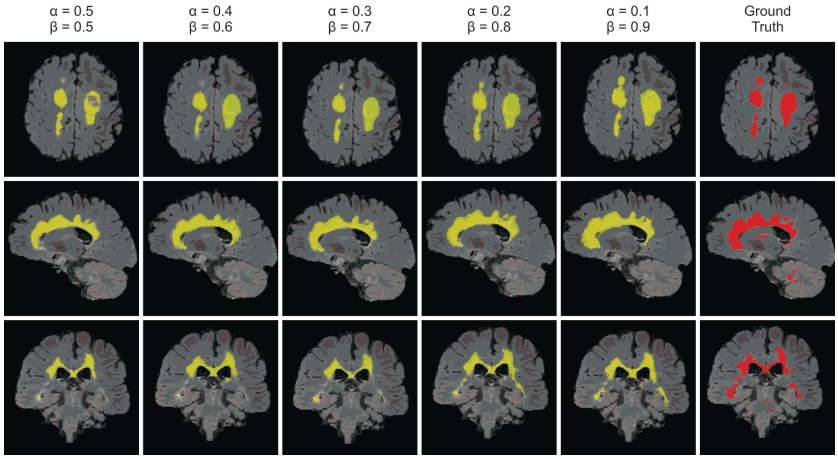

Fig. 3. The effect of different penalties on FP and FN in the Tverskey loss function on a case with extremely high density of lesions. The best results were obtained at $\beta = 0.7$

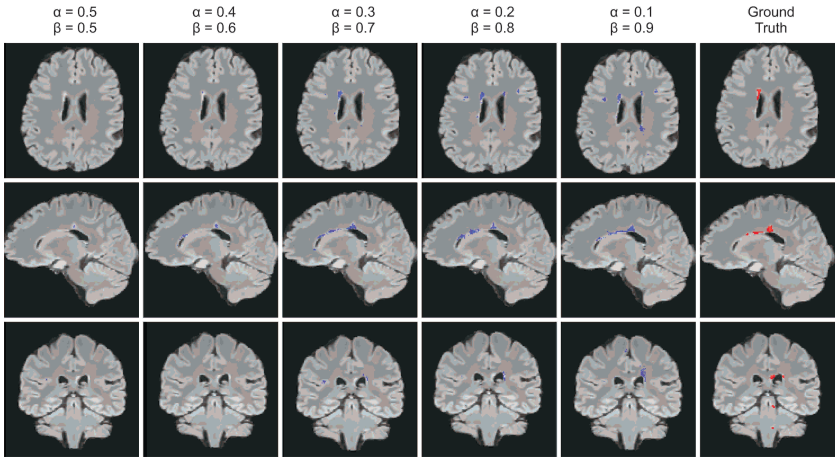

Fig. 4. The effect of different penalties on FP and FN in the Tverskey loss function on a case with extremely low density of lesions. The best results were obtained at $\beta = 0.7$.

4 Discussion and conclusion

We introduced a new loss function based on the Tversky index, that generalizes the Dice coefficient and F_β scores, to achieve improved trade-off between precision and recall in segmenting highly unbalanced data via deep learning. To this end, we added our proposed loss layer to a state-of-the-art 3D fully convolutional deep neural network based on the U-net architecture [16,17]. Experimental results in MS lesion segmentation show that all performance evaluation metrics (on the test data) improved by using the Tversky loss function rather than using the Dice similarity coefficient in the loss layer. While the loss function was deliberately designed to weigh recall higher than precision (at $\beta = 0.7$), consistent improvements in all test performance metrics including DSC and F_2 scores on the test set indicate improved generalization through this type of training. Compared to DSC which weighs recall and precision equally, and the ROC analysis, we consider the area under the PR curves (APR, shown in Figure 2) the most reliable performance metric for such highly skewed data [8,1]. To put the work in context, we reported average DSC, F_2, and APR scores (equal to 56.4, 57.3, and 56.0, respectively), which indicate that our approach performed very well compared to the latest results in MS lesion segmentation [6,20]. We did not conduct a direct comparison in the application domain, however, as this paper intended to provide proof-of-concept on the effect and usefulness of the Tverky loss layer (and F_β scores) in deep learning. Future work involves training and testing on larger, standard datasets in multiple applications to compare against state-of-the-art segmentations using appropriate performance criteria.

Acknowledgements

This work was in part supported by the National Institutes of Health (NIH) under grant R01 EB018988. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Boyd, K., Eng, K.H., Page, C.D.: Area under the precision-recall curve: Point estimates and confidence intervals. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 451–466. Springer (2013)
2. Brosch, T., Yoo, Y., Tang, L.Y., Li, D.K., Traboulsee, A., Tam, R.: Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–11. Springer (2015)
3. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. NeuroImage (2017)
4. 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. pp. 424–432. Springer (2016)
5. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 411–418. Springer (2013)
6. Commowick, O., Cervenansky, F., Ameli, R.: Msseg challenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure (2016)
7. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240. ACM (2006)
8. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters 27(8), 861–874 (2006)
9. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis 35, 18–31 (2017)
10. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78 (2017)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on. pp. 565–571. IEEE (2016)
14. Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I.: Automatic segmentation of mr brain images with a convolutional neural network. IEEE transactions on medical imaging 35(5), 1252–1261 (2016)

15. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in mri images. IEEE transactions on medical imaging 35(5), 1240–1251 (2016)
16. 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. Springer (2015)
17. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Auto-context convolutional neural network for geometry-independent brain extraction in magnetic resonance imaging. arXiv preprint arXiv:1703.02083 (2017)
18. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 39(4), 640–651 (2017)
19. Tversky, A.: Features of similarity. Psychological review 84(4), 327 (1977)
20. Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X.: Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. NeuroImage 155, 159–168 (2017)
21. Wachinger, C., Reuter, M., Klein, T.: Deepnat: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage (2017)
22. Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 108, 214–224 (2015)