TopK Dice Loss for Medical Image Segmentation

Seyed Mohsen Hosseini
smhosseini741@gmail.com
University of Tehran, Iran

Abstract

In image segmentation tasks, the class-imbalance problem affects the performance of neural networks. This problem is especially severe in medical images where usually the task involves segmenting a small foreground in a large background. Different methods have been developed to address this problem, e.g. using region based losses like the Dice loss. There is also a difficulty-imbalance problem, because the majority of samples of each class are easy to classify and hard samples are in minority. This leads to an ineffective learning, as the easy samples dominate the training process. A common strategy to address this problem, is reweighting the cross entropy loss, in order to focus the training on hard samples. A novel loss for addressing both the class-imbalance and difficulty-imbalance problems is proposed in this work. Unlike previous methods that address these problems separately, the proposed method tackles them at the same time using a modified Dice loss. Experiments on different medical segmentation tasks show that the proposed method outperforms popular existing methods for addressing class and difficulty imbalance problems.

1 Introduction

When there is an extreme imbalance between the numbers of examples of each class in a dataset it is considered imbalanced. The effect of Imbalanced data on the performance of deep neural networks in different tasks is a well-studied area [20, 17, 28]. To minimize the CE loss, neural networks minimize the loss of the classes that make up the majority of the loss, this means the network is biased towards the classes with most samples and the importance of each class is determined by the number of its samples. This will lead to poor performance for under-represented classes.

In segmentation tasks where a large number of classifications should be done on pixels of an image, the imbalance between classes can get very large, as the size of the objects can be very small compared to background. In some tasks the ratio between background and foreground pixels can be more that 1,000 to 1 [1]. The obvious solution would be to use datasets with uniform distribution of classes, but in practice producing such datasets is difficult and costly especially in medical imaging tasks.

Different methods have been developed to improve the performance of the network for underrepresented classes when a network is trained with imbalance data. There are mainly three categories of methods: data re-sampling, using region based losses, and weighting the loss values. Re-sampling addresses the imbalance problem at the data level. In these methods, the under-represented classes are repeated and part of the over-represented classes is removed from the training data. Increasing the training time, removing useful information from the dataset, and increasing the probability of over-fitting, are among the drawbacks of these methods [28]. In [21], images containing rare or finely annotated objects are duplicated in the training data and using this technique they report improved segmentations in hard classes like bicycle.

Using region based losses, like the Dice loss, is a popular method to improve the performance of a model trained with imbalanced data in medical segmentation tasks. A review of different loss functions for medical segmentation can be found here [1]. In region based losses, instead of producing independent losses for each pixel of the image, a loss is calculated based on the mismatch of the segmented area and ground truth. A misclassification of a small object has a small effect on the overall cross entropy loss, but it creates a large Dice loss because the mismatch between the predicted and the ground truth areas would be high and it is not related to the size of the object. This property of Dice loss and also the fact that Dice similarity score is one of the main metrics for network evaluation have made the Dice loss a popular loss in medical segmentation tasks.

In generalized Dice loss [22], the inverse of class volume is used to balance the effect of classes with different sizes on the overall loss. A region loss with more control over the impact of false positives and false negatives on the loss is offered as Tversky loss [23]. Because of the sensitivity of Dice loss towards smaller objects [5], small mislabelling in ground truth can result in large gradients which make the training more difficult. To address this problem [15] suggests calculating the Dice loss for the whole mini-batch, instead of averaging individual Dice losses of samples in a mini-batch.

Using a combination of different losses for imbalanced data is a common practice [1, 24]. In the combination of CE loss and Dice loss [15, 16, 25], while Dice loss provides more sensitivity for smaller objects and improves the performance for imbalanced data, CE loss leads to smoother gradient [25]. The combination of Dice and CE loss performs better than Dice loss alone [1]. Weighted cross entropy is another group of methods for solving the class imbalanced problem. In these methods different weights are applied to different classes and the majority classes are down-weighted.

In [19], the weight is calculated based on the size of a lesion as well as its class, so lesions of the same class receive different weights based on their size and higher weights are given to smaller lesions. Weighted cross entropy when the weights assigned to each class are based on the inverse of class frequency is also called balanced cross entropy [4, 17]. The performance of balanced CE in medical segmentation tasks is generally lower than CE [1, 5], and therefore it is not commonly used in popular medical segmentation tasks. Balance CE produces lower false negative but higher false positives and it is not a very effective method to address the class imbalance problem [5].

Difficulty imbalance is another imbalance in training data. The majority of pixels of an image are easily classified and produce low losses and usually a small part of the image is where most of the errors happen. This imbalance is related to the class imbalance, as the minority classes are usually harder to classify. But difficult to classify areas can also exist in majority classes, and by down-weighting the majority classes those areas are also down-weighted. Therefore methods addressing the class imbalance problem do not automatically address the difficulty imbalance problem.

In existing methods for addressing the difficulty imbalance problem, hard to classify pixels are assigned higher weights. The weights are usually determined based on predicted probabilities [2, 3]. Another reweighting scheme is based on the position of pixels in the image. Higher weights are applied to areas of the image where error is expected to be high, like boundary areas. In the original Unet [4], higher weights are given to classes based on their frequency, and also in order to have a more accurate segmentation around the boundaries of the cells, those areas are given higher weights to increase their impact on the loss. In the task of edge detection where the edge pixels are the minority class, the inverse of their frequency is used in [18] to determine their balancing weight.

In focal cross entropy loss [2], the cross entropy loss of pixels are modulated based on their predicted probabilities. When the predicted probability of a misclassified pixel is close to zero the modulating value is close to one and the loss is not reduced very much. For well-classified pixels the modulating value is close to zero reducing the loss. This method allows the network to focus more on hard to classify examples. In [27], pixels with losses lower than a certain threshold are ignored. For each mini-batch if the performance is high the loss threshold is increased otherwise it is lowered. In topK cross entropy loss [3], difficulty as well as the number of hard samples are considered. Pixels are ranked based on their classification error and only k percent of pixels from the top of that list are kept and the rest are ignored.

The minority classes are more likely to be misclassified but because the data is mostly populated by the majority classes, the misclassified pixels are mostly from the majority classes. Therefore by focusing the training on difficult pixels, and not considering their class, the bias towards the majority class will not be addressed. So methods that improve class imbalance problem don't address the difficulty imbalance problem and vice versa. Different methods have been developed to tackle the class and difficulty imbalance problems at the same time. In [26], for the tasks of image classification and verb classification in video clips, difficulty and class are considered together to determine the balancing weights. The weights are calculated based on the classification difficulty of each class using validation accuracy of that class after each epoch.

In segmentation tasks a combination of reweighted CE loss and Dice loss is usually used to handle the class and difficulty imbalance at the same time. A reweighted CE loss for difficulty imbalance is combined with the Dice loss for class imbalance. DiceFocal (focal CE + Dice loss) [1] and DiceTopK (topK CE + Dice loss) [29] are the two commonly implemented compound losses for addressing the two types of imbalance in the data. But most of the top performing models just use CE + Dice loss [15, 16, 25], which shows that focal and topK losses are not very effective in handling the difficulty imbalance.

In the proposed method, instead of using two losses, a modified Dice loss is used to handle both types of imbalance. The Dice loss is very effective in correcting the class imbalance in medical segmentation tasks. The main motivation behind the proposed method is to try to take advantage of the Dice loss to also handle difficulty imbalance. In existing methods a variation of weighted CE loss is utilised to tackle the difficulty imbalance. Compared to Dice loss weighted CE is not very effective in handling class imbalance [5]. And as the results of this study show, Dice loss is also more effective than weighted CE for improving the difficulty imbalance problem. Experiments on three popular medical segmentation tasks are conducted, and the results show that the proposed method outperforms the existing class and difficulty balancing methods.

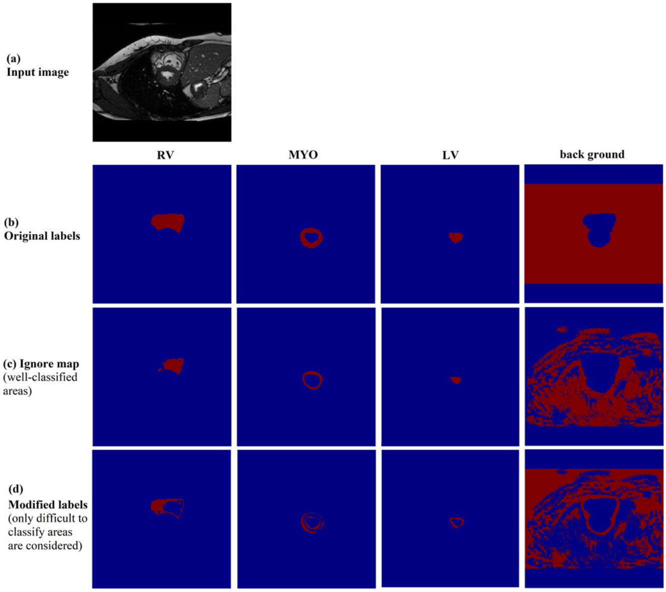

Figure 1: Modified ground truth labels for topK Dice loss calculation. (a) Input image, (b) original ground truth labels, (c) Ignore map (well classified areas that will be excluded from the Dice loss calculation), (d) modified ground truth (well classified areas are set to zero).

2 Proposed Method

The proposed method is based on focusing the Dice loss on the areas of the image that are predicted with lower accuracies. Similar to topK CE loss, k percent of the pixels with highest error, in each class, are kept and rest are ignored. The ignored areas are not included in Dice loss calculations. After identifying the well-classified areas, based on the predicted probabilities, those areas are set to zero in ground truth labels and predicted probability maps, producing their modified versions. Figure 1 depicts an example of this process. The squared version of the Dice loss is defined as:

$$Dice\ loss = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i=1}^{H}\sum_{j=1}^{W} Y_{ij}^{c}\, P_{ij}^{c}}{\sum_{i=1}^{H}\sum_{j=1}^{W}(Y_{ij}^{c})^2 + \sum_{i=1}^{H}\sum_{j=1}^{W}(P_{ij}^{c})^2} \qquad (1)$$

Where Y^c, and P^c are the ground truth label and predicted probability maps. H and W, are the image dimensions. C is the number of classes that are present in the image.

In order to find the well classified pixels, the hadamard product of the predicted maps and ground truth labels is calculated, producing the true positive map of each class, as:

$$TP^c = Y^c \odot (P^c + \mathcal{E}) \qquad (2)$$

Easy to classify pixels have higher values in TP^c. In order to keep K percent of most difficult pixels of each class, the values of TP^c are sorted from high to low and the D^c-th highest value in the list is chosen as the difficulty threshold: t^c. This threshold is used to exclude the well classified areas of the image from the Dice loss calculations. D^c is $(100 - K)$ percent of the total number of foreground pixels in each class, equation (3). K percent of foreground pixels in each class are kept and $(100 - K)$ percent are excluded.

Because the numbers are usually rounded in the results of the network, a small random perturbation (Ɛ) is added to the true positive map in equation (2). Without Ɛ, pixels with the same accuracy may be excluded from or included in the well classified areas, causing the number of pixels in those areas to be different from the K percent desired values. Ɛ is generated once and is not changed during training and is the same for all classes.

$$D^c = (1 - K/100) \times \sum_{i=1}^{H} \sum_{j=1}^{W} Y_{ij}^c \qquad (3)$$

Areas of the image that will be ignored in the topK Dice loss calculation for each class are defined as:

$$Ignore\_map^c = \mathbf{1}\{TP^c > t^c\} \qquad (4)$$

Where $\mathbf{1}\{x\} = 1$, if x is true. Easy to classify pixels with probabilities higher than t^c are ignored. Using $Ignore\_map^c$ (Figure 1 (c)) well classified pixels are set to zero in ground truth labels and predicted probabilities:

$$Yk^c = Y^c \odot (1 - Ignore\_map^c) \qquad (5)$$
$$Pk^c = P^c \odot (1 - Ignore\_map^c) \qquad (6)$$

And finally the topK Dice loss is calculated based on the modified ground truth and modified network predictions:

$$topK\ Dice\ loss = 1 - \frac{1}{C} \sum_{c=1}^{C} \frac{2 \sum_{i=1}^{H} \sum_{j=1}^{W} Yk_{ij}^c \ Pk_{ij}^c}{\sum_{i=1}^{H} \sum_{j=1}^{W} (Yk_{ij}^c)^2 + \sum_{i=1}^{H} \sum_{j=1}^{W} (Pk_{ij}^c)^2} \qquad (7)$$

In topK CE and focal CE the importance of difficult pixels are increased by giving them a higher weight. But in topK Dice loss the importance of difficult pixels is increased by reducing the number of pixels that are included in the Dice calculation. This happens by taking advantage of the sensitivity of Dice score towards smaller objects [5]. For large objects, small numbers of false negatives do not make a significant difference in the total number of true positives (numerator in equation (8)), and Dice score is not affected too much. But small errors in a small object cause more significant reduction in true positives and thus lower the Dice score. So errors in difficult to classify pixels which are usually the details of the predicted objects do not create a large Dice loss, especially in larger objects.

$$Dice\ loss = 1 - Dice\ score = 1 - \frac{TP}{TP + 0.5 \times FP + 0.5 \times FN} \qquad (8)$$

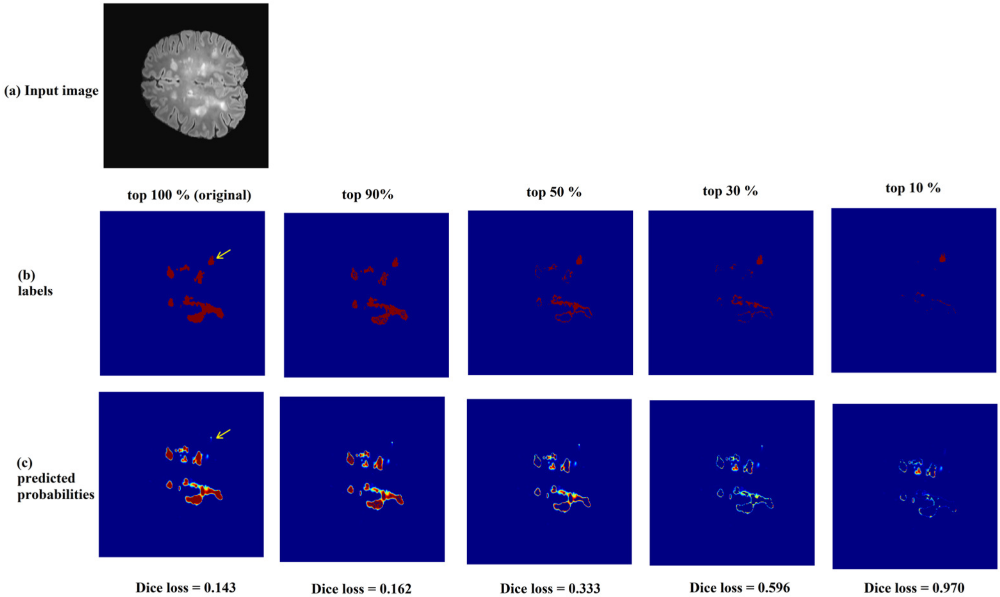

Figure 2: The effect of concentrating on difficult to classify pixels on the Dice loss. The Dice loss is only calculated for the top K% (K = 100, 90, 50, 30, 10) pixels of the image that are hardest for the network to classify.

By utilising only a part of the original ground truth labels, we decrease the number of true positives and increase the sensitivity of the Dice score towards false negatives. In Figure 2 it can be seen that in case of the original label the Dice loss is low even though a part of the label is completely missed by the model (indicated by the arrow). This is because the missed areas are a small percentage of the fore ground. The effect of decreasing the value of K can be seen in this Figure. The Dice loss increases with lower k values, as the modified ground truth labels concentrate more on the high error areas, like the boundary areas and also the section completely missed by the model. The label in the case of top 10% is mostly on that section.

Making the Dice loss more sensitive towards false negatives is also addressed in Tversky loss [23]. Two hyper parameters, α and β, are set as the coefficients of false positives (FP) and false negatives (FN) respectively. To put more emphases on FN the hyper parameters are chosen as, $\alpha = 0.3$, and $\beta = 0.7$. These constant parameters increase the impact of FN on the Dice loss, leading to lower FN values. As there is no mechanism to focus on difficult pixels, the well classified pixels, which are in the majority, have the most effect on the reduction of FN, while the minority difficult pixels have limited effect. This loss produces lower performances compared to topK CE + Dice loss, and Focal CE + Dice loss [1].

3 Experiments

Multiple experiments on popular medical segmentation tasks are conducted to evaluate the performance of topK Dice loss. In all experiments, to have a more meaningful comparison between the loss functions, the training conditions, e.g. the network, training data, augmentation, learning rate, are kept the same with the only different factor being the loss function.

3.1 Datasets

Synapse multi-organ dataset[1] from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge contains 30 abdominal scans with 3779 slices. The number of slices in each scan varies from 85 to 198 slices with 512 × 512 resolution. Following [11, 12], 18 scans are used for training and 12 scans are used for testing. The data split is the same as [12], and following [11,12], only 8 organs are segmented in this task including aorta, gallbladder, right kidney, left kidney, liver, pancreas, spleen and stomach. For training and testing the input slices are resized to 320 × 320 and their values are clipped within values -125 and 275 and then normalized to [0, 1]. Data augmentations applied to the training data are random flipping, rotation, zooming, elastic deformation, and Gaussian blurring. For this task, the network is trained for 300 epochs with different losses.

The Automated Cardiac Diagnosis dataset2 (ACDC) contains cardiac 3D MRI scans of 150 patients. The task is to segment left ventricles (LV), right ventricles (RV), and the myocardium (MYO). Following [12], out of the 100 scans of the training set, 70 and 20 scans are randomly chosen for training and testing. For training and testing the input slices are resized to 256 × 256 and normalized to [0, 1]. Data augmentations for this task are random flipping, rotation, and zooming, elastic deformation, and the network is trained for 300 epochs with different losses.

The MSSEG dataset[3] consists of 15 brain scans, and the task is the binary segmentation of multiple sclerosis (MS) lesions. The ground truth labels are the consensus between 4 different expert annotations. 9 scans were chosen randomly as training set and the rest used as test set. All 5 available modalities are used as the input of the network. For training and testing the input slices are resized to 256 × 256 and normalized to [0, 1]. Data augmentations are rotation and zooming. For this task, the network is trained for 100 epochs using different losses.

3.2 Implementation

The implemented network, for Synapse and ACDC datasets, is a U-net [5] based on Resnet34 [6]. The network has 5 max pooling layers to down-sample to 1/32 of the original resolution. Deep supervision is utilized for training. Segmentation heads were added to the network at scales: 1/4, 1/8 and 1/16. A segmentation head consists of a convolution layer, bilinear up-sampling to the original resolution and a soft-max layer. The final loss of the network is the sum of losses calculated from different scales. Like many previous studies [13, 14], the output of the network used to produce the segmentation at inference time is the up-scaled prediction map from the 1/4 scale. For MSSEG dataset, because the images contain more details, a simple U-net [5] with 4 max pooling layers (1/16 of the original resolution) is implemented. Segmentation heads were added to the network at scales: 1, 1/2 and 1/4. For both networks, Weight decay of 1e-4 and the Adam optimizer with initial learning rate of 5e-4 are used. The learning rate is reduced according to a polynomial learning rate policy,

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/

[3] https://portal.fli-iam.irisa.fr/msseg-2/

Loss functions	mDice	mIoU	Aorta	Gallbladder	Kidney left	Kidney right	Liver	Pancreas	Spleen	Stomach
CE	81.10	72.30	90.50	60.65	83.77	81.23	95.38	64.94	91.71	80.66
CE + Dice	81.75	72.74	91.79	60.48	85.54	83.67	95.49	68.08	89.79	79.19
focal + Dice	81.26	72.26	91.02	59.10	87.35	84.77	95.31	64.98	90.30	77.27
topK + Dice	81.87	73.00	91.58	61.73	85.25	82.80	95.86	65.63	90.80	81.32
CE + topK-Dice (our)	82.73	74.07	91.85	61.16	88.78	84.72	95.34	68.75	92.04	79.18

Table 1: Results of the Synapse multi-organ dataset. The dice score of 8 organs are also reported. The best and second best results are marked in red and blue respectively.

Loss functions	mDice	mIoU	RV	MYO	LV
CE	91.29	84.42	90.07	88.91	94.89
CE + Dice	91.29	84.43	90.37	89.05	94.45
Focal-CE + Dice	91.24	84.39	89.45	88.82	95.45
topK-CE + Dice	91.34	84.53	90.04	89.04	94.93
CE + topK-Dice (our)	91.67	85.03	90.20	89.57	95.24

Table 2: Results of the ACDC multi-organ dataset. The dice score of 3 organs are also reported. The best and second best results are marked in red and blue respectively.

Loss functions	mDice	mIoU	mPrec.	mRec.
CE	65.37	48.78	76.72	58.82
CE + Dice	66.29	49.94	73.37	61.95
Focal-CE + Dice	66.31	50.31	68.86	67.83
topK-CE + Dice	65.93	49.46	73.45	61.75
CE + topK-Dice (our)	67.19	51.10	71.69	66.03

Table 3: Results of the MSSEG dataset. The best and second best results are marked in red and blue respectively.

$(\text{initial learning rate } \times (1 - \text{iter.}/\text{max\_iter.})^{0.9})$. The value of K used in topK Dice loss is set empirically to 50% for Synapse and ACDC datasets, and 30% for MSSEG dataset. The K is chosen the same for topK CE loss. A CE loss is also added to topK Dice loss. This is to make the gradient smoother [25], which is helpful because the topK Dice loss is concentrated on difficult pixels and smaller areas of the image. The final loss is the summation of CE and topK Dice loss. The proposed loss is compare with 4 popular losses.

The evaluation metrics are mean Dice score (mDCS), mean IoU (mIoU), mean precision (mPrec.) and mean recall (mRec.) which are averaged over the test set. In Synapse and ACDC datasets, DCS and IoU of each organ is averaged over the test scans and then averaged over different organs to produce the overall metrics.

3.3 Results

The qualitative comparison results on three different datasets are presented in this section. As it can be seen in tables 1 to 3 the performance of the proposed method is higher than topK CE + Dice loss, and focal CE + Dice loss, which are the most common methods for handling the class and difficulty imbalance problems. This shows that topK-Dice loss is more effective than existing common methods in handling the difficulty imbalance, in binary as well as multiclass segmentation tasks. Figures 3 and 4 show visual comparison of the results for diffident datasets.

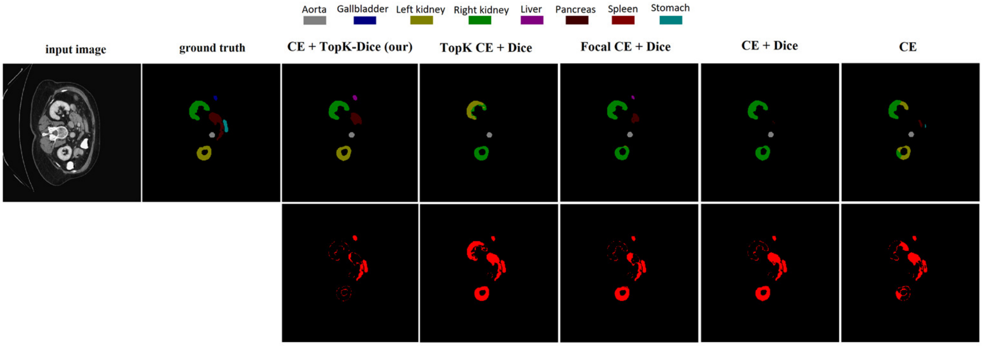

Figure 3: visual comparison of the results from Synapse multi-organ segmentation task. The proposed TopK-Dice loss is compared with commonly used loss functions. The first row contains the label predictions and the second row shows the error map (incorrectly labeled pixels are marked by red).

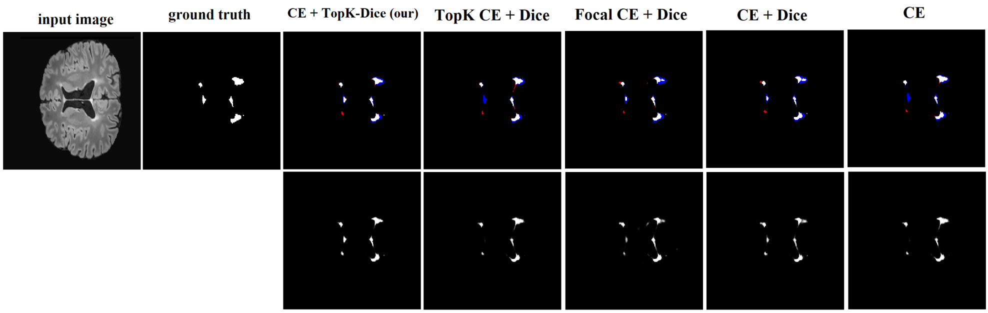

Figure 4: visual comparison of the results from MSSEG segmentation task. The proposed TopK-Dice loss is compared with commonly used loss functions. The first row contains the label predictions (false positives and false negatives are marked by red and blue) and the second row shows the predicted probabilities.

4 Conclusion

A novel modified Dice loss for handling both class and difficulty imbalance problems, is proposed in this study. The proposed loss is different from the common strategy, which is to use the sum of a reweighed CE loss (focal CE or topK CE) for difficulty imbalance, and the Dice loss for class imbalance. The proposed topK Dice loss outperforms the reweighting CE strategies in different medical segmentation tasks. This shows that, the topK Dice loss can successfully handle both class and difficulty imbalance problems at the same time, utilising a modified Dice loss.

References

[1] Ma, Jun, et al. "Loss odyssey in medical image segmentation." Medical Image Analysis 71 (2021): 102035.

[2] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[3] Pohlen, Tobias, et al. "Full-resolution residual networks for semantic segmentation in street scenes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[4] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[5] Bertels, Jeroen, et al. "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice." Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. Springer International Publishing, 2019.

[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Fitzgerald, Kerr, and Bogdan Matuszewski. "FCB-SwinV2 Transformer for Polyp Segmentation." arXiv preprint arXiv:2302.01027 (2023).

[8] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen & H.D. Johansen , "Kvasir-SEG: A Segmented Polyp Dataset," in International Conference on Multimedia Modeling, 2019.

[9] J. Bernal, F. J. Sanchez, G. Fern ´andez-Esparrach, D. Gil, C. Rodríguez & F. Vilariño,, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," Computerized Medical Imaging and Graphics, vol. 43, pp. 99-111, 2015.

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368, 2019.

[11] Fu, S., Lu, Y., Wang, Y., Zhou, Y., Shen, W., Fishman, E., Yuille, A.: Domain adaptive relational reasoning for 3d multi-organ segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656-666. Springer (2020).

[12] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

[13] Fan, Deng-Ping, et al. "Pranet: Parallel reverse attention network for polyp segmentation." International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2020.

[14] Kim, Taehun, Hyemin Lee, and Daijin Kim. "Uacanet: Uncertainty augmented context attention for polyp segmentation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[15] Isensee, Fabian, et al. "nnU-Net for brain tumor segmentation." Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in

Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6. Springer International Publishing, 2021.

[16] Rahman, Md Mostafijur, and Radu Marculescu. "Multi-scale Hierarchical Vision Transformer with Cascaded Attention Decoding for Medical Image Segmentation." arXiv preprint arXiv:2303.16892 (2023).

[17] Cui, Yin, et al. "Class-balanced loss based on effective number of samples." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[18] Holistically-nested edge detection. In Proceedings of the IEEE international conference on computer vision, pages 1395–1403, 2015.

[19] Shirokikh, Boris, et al. "Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation." Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer International Publishing, 2020.

[20] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[21] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

[22] Sudre, Carole H., et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations." Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. Springer International Publishing, 2017.

[23] Salehi, Seyed Sadegh Mohseni, Deniz Erdogmus, and Ali Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks." International workshop on machine learning in medical imaging. Cham: Springer International Publishing, 2017.

[24] Taghanaki, Saeid Asgari, et al. "Combo loss: Handling input and output imbalance in multi-organ segmentation." Computerized Medical Imaging and Graphics 75 (2019): 24-33.

[25] Zhou, Zongwei, et al. "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation." IEEE transactions on medical imaging 39.6 (2019): 1856-1867.

[26] Sinha, Saptarshi, Hiroki Ohashi, and Katsuyuki Nakamura. "Class-wise difficulty-balanced loss for solving class-imbalance." Proceedings of the Asian conference on computer vision. 2020.

[27] Wu, Zifeng, Chunhua Shen, and Anton van den Hengel. "Bridging category-level and instance-level semantic image segmentation." arXiv preprint arXiv:1605.06885 (2016).

[28] Huang, Chen, et al. "Learning deep representation for imbalanced classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[29] Brugnara, Gianluca, Fabian Isensee, Ulf Neuberger, David Bonekamp, Jens Petersen, Ricarda Diem, Brigitte Wildemann et al. "Automated volumetric assessment with artificial neural networks might enable a more accurate assessment of disease burden in patients with multiple sclerosis." European radiology 30:2356-2364, 2020.