Optimization with Soft Dice Can Lead to a Volumetric Bias

Jeroen Bertels[1(✉)] , David Robben[1,2], Dirk Vandermeulen[1],
and Paul Suetens[1]

[1] Processing Speech and Images, ESAT, KU Leuven, Leuven, Belgium
`jeroen.bertels@kuleuven.be`
[2] icometrix, Leuven, Belgium

Abstract. Segmentation is a fundamental task in medical image analysis. The clinical interest is often to measure the volume of a structure. To evaluate and compare segmentation methods, the similarity between a segmentation and a predefined ground truth is measured using metrics such as the Dice score. Recent segmentation methods based on convolutional neural networks use a differentiable surrogate of the Dice score, such as soft Dice, explicitly as the loss function during the learning phase. Even though this approach leads to improved Dice scores, we find that, both theoretically and empirically on four medical tasks, it can introduce a volumetric bias for tasks with high inherent uncertainty. As such, this may limit the method's clinical applicability.

Keywords: Segmentation · Cross-entropy · Soft Dice · Volume

1 Introduction

Automatic segmentation of structures is a fundamental task in medical image analysis. Segmentations either serve as an intermediate step in a more elaborate pipeline or as an end goal by itself. The clinical interest often lies in the volume of a certain structure (e.g. the volume of a tumor, the volume of a stroke lesion), which can be derived from its segmentation [11]. The segmentation task can also carry inherent uncertainty (e.g. noise, lack of contrast, artifacts, incomplete information).

To evaluate and compare the quality of a segmentation, the similarity between the true segmentation (i.e. the segmentation derived from an expert's delineation of the structure) and the predicted segmentation must be measured. For this purpose, multiple metrics exist. Among others, overlap measures (e.g. Dice score, Jaccard index) and surface distances (e.g. Haussdorf distance, average surface distance) are commonly used [13].

The focus on one particular metric, the Dice score, has led to the adoption of a differentiable surrogate loss, the so-called soft Dice [9,15,16], to train convolutional neural networks (CNNs). Many state-of-the-art methods clearly outperform the established cross-entropy losses using soft Dice as loss function [7,12].

J. Bertels and D. Robben—Contributed equally to this work.

In this work, we investigate the effect on volume estimation when optimizing a CNN w.r.t. cross-entropy or soft Dice, and relate this to the inherent uncertainty in a task. First, we look into this volumetric bias theoretically, with some numerical examples. We find that the use of soft Dice leads to a systematic under- or overestimation of the predicted volume of a structure, which is dependent on the inherent uncertainty that is present in the task. Second, we empirically validate these results on four medical tasks: two tasks with relatively low inherent uncertainty (i.e. the segmentation of third molars from dental radiographs [8], BRATS 2018 [4–6,14]) and two tasks with relatively high inherent uncertainty (i.e. ISLES 2017 [2,18], ISLES 2018 [3]).

2   Theoretical Analysis

Let us formalize an image into I voxels, each voxel corresponding to a true class label c_i with $i = 0 \ldots I - 1$, forming the true class label map $C = [c_i]^I$. Typical in medical image analysis, is the uncertainty of the true class label map C (e.g. due to intra- and inter-rater variability; see Sect. 2.2). Under the assumption of binary image segmentation with $c_i \in \{0, 1\}$, a probabilistic label map can be constructed as $Y = [y_i]^I$, where each $y_i = P(c_i = 1)$ is the probability of y_i belonging to the structure of interest. Similarly, we have the maps of voxel-wise label predictions $\hat{C} = [\hat{c}_i]^I$ and probabilities $\hat{Y} = [\hat{y}_i]^I$. In this setting, the class label map $\hat{C}$ is constructed from the map of predictions $\hat{Y}$ according to the highest likelihood.

The Dice score $\mathcal{D}$ is defined on the label maps as:

$$\mathcal{D}(C, \hat{C}) = \frac{2|C \cap \hat{C}|}{|C| + |\hat{C}|} \tag{1}$$

The volumes $\mathcal{V}(C)$ of the true structure and $\mathcal{V}(\hat{C})$ of the predicted structure are then, with v the volume of a single voxel:

$$\mathcal{V}(C) = v \sum_{i=0}^{I-1} c_i, \ \ \mathcal{V}(\hat{C}) = v \sum_{i=0}^{I-1} \hat{c}_i \tag{2}$$

In case the label map is probabilistic, we need to work out the expectations:

$$\mathcal{V}(Y) = v\mathbf{E}[\sum_{i=0}^{I-1} y_i], \ \ \mathcal{V}(\hat{Y}) = v\mathbf{E}[\sum_{i=0}^{I-1} \hat{y}_i] \tag{3}$$

2.1   Risk Minimization

In the setting of supervised and gradient-based training of CNNs [10] we are performing empirical risk minimization. Assume the CNN, with a certain topology,

is parametrized by $\boldsymbol{\theta} \in \Theta$ and represents the functions $\mathcal{H} = \{\mathfrak{h}_{\boldsymbol{\theta}}\}^{|\Theta|}$. Further assume we have access to the entire joint probability distribution $P(\mathbf{x}, y)$ at both training and testing time, with $\mathbf{x}$ the information (for CNNs this is typically a centered image patch around the location of y) of the network that is used to make a prediction $\hat{y} = \mathfrak{h}_{\boldsymbol{\theta}}(\mathbf{x})$ for y. For these conditions, the general risk minimization principle is applicable and states that in order to optimize the performance for a certain non-negative and real-valued loss $\mathcal{L}$ (e.g. the metric or its surrogate loss) at test time, we can optimize the same loss during the learning phase [17]. The risk $\mathcal{R}_{\mathcal{L}}(\mathfrak{h}_{\boldsymbol{\theta}})$ associated with the loss $\mathcal{L}$ and parametrization $\boldsymbol{\theta}$ of the CNN, without regularization, is defined as the expectation of the loss function:

$$\mathcal{R}_{\mathcal{L}}(\mathfrak{h}_{\boldsymbol{\theta}}) = \mathbf{E}[\mathcal{L}(\mathfrak{h}_{\boldsymbol{\theta}}(\mathbf{x}), y)] \tag{4}$$

For years, minimizing the negative log-likelihood has been the gold standard in terms of risk minimization. For this purpose, and due to its elegant mathematical properties, the voxel-wise cross-entropy loss ($\mathcal{CE}$) is used:

$$\mathcal{CE}(\hat{Y}, Y) = \sum_{i=0}^{I-1} [\mathcal{CE}(\hat{y}_i, y_i)] = -\sum_{i=0}^{I-1} [y_i \log \hat{y}_i] \tag{5}$$

More recently, the soft Dice loss ($\mathcal{SD}$) is used in the optimization of CNNs to directly optimize the Dice score at test time [9,15,16]. Rewriting Eq. 1 to its non-negative and real-valued surrogate loss function as in [9]:

$$\mathcal{SD}(\hat{Y}, Y) = 1 - \frac{2 \sum_{i=0}^{I-1} \hat{y}_i y_i}{\sum_{i=0}^{I-1} \hat{y}_i + \sum_{i=0}^{I-1} y_i} \tag{6}$$

2.2   Uncertainty

There is considerable uncertainty in the segmentation of medical images. Images might lack contrast, contain artifacts, be noisy or incomplete regarding the necessary information (e.g. in ISLES 2017 we need to predict the infarction after treatment from images taken before, which is straightforwardly introducing *inherent* uncertainty). Even at the level of the true segmentation, uncertainty exists due to intra- and inter-rater variability. We will investigate what happens with the estimated volume $\mathcal{V}$ of a certain structure in an image under the assumption of having perfect segmentation algorithms (i.e. the prediction is the one that minimizes the empirical risk).

Assuming independent voxels, or that we can simplify Eq. 3 into J independent regions with true uncertainty p_j and predicted uncertainty $\hat{p}_j$, and corresponding volumes $s_j = vn_j$, with n_j the number of voxels belonging to region $j = 0 \ldots J - 1$ (having each voxel as an independent region when $n_j = 1$), we get:

$$\mathcal{V}(Y) = \sum_{j=0}^{J-1} (s_j p_j), \ \mathcal{V}(\hat{Y}) = \sum_{j=0}^{J-1} (s_j \hat{p}_j) \tag{7}$$

We analyze for $\mathcal{CE}$ the predicted uncertainty that minimizes the risk $\mathcal{R}_{\mathcal{CE}}(\mathfrak{h}_{\boldsymbol{\theta}})$:

$$\arg\min_{\hat{Y}}[\mathcal{R}_{\mathcal{CE}}(\mathfrak{h}_{\boldsymbol{\theta}})] = \arg\min_{\hat{Y}}[\mathbf{E}[\mathcal{CE}(\hat{Y}, Y)]] \tag{8}$$

We need to find for each independent region j:

$$\arg\min_{\hat{p}_j}[s_j\mathcal{CE}(\hat{p}_j, p_j)] = \arg\min_{\hat{p}_j}[-p_j\log\hat{p}_j - (1-p_j)\log(1-\hat{p}_j)] \tag{9}$$

This function is continuous and its first derivative monotonously increasing in the interval $]0,1[$. First order conditions w.r.t. $\hat{p}_j$ give the optimal value for the predicted uncertainty $\hat{p}_j = p_j$. With the predicted uncertainty being the true uncertainty, $\mathcal{CE}$ becomes an unbiased volume estimator.

We analyze for SD the predicted uncertainty that minimizes the risk $\mathcal{R}_{\mathcal{SD}}(\mathfrak{h}_{\boldsymbol{\theta}})$:

$$\arg\min_{\hat{Y}}[\mathcal{R}_{\mathcal{SD}}(\mathfrak{h}_{\boldsymbol{\theta}})] = \arg\min_{\hat{Y}}[\mathbf{E}[SD(\hat{Y}, Y)]] \tag{10}$$

We need to find for each independent region j:

$$\arg\min_{\hat{Y}}[\mathbf{E}[SD(\hat{Y}, Y)]] = \arg\min_{\hat{p}_j}[\mathbf{E}[1 - \frac{2\sum_{j=0}^{J-1}s_j\hat{p}_jp_j}{\sum_{j=0}^{J-1}s_j\hat{p}_j + \sum_{j=0}^{J-1}s_jp_j}]] \tag{11}$$

This minimization is more complex and we analyze its behavior by inspecting the values of SD numerically. We will consider the scenarios with only a single region or with multiple independent regions with inherent uncertainty in the image. For each scenario we will vary the inherent uncertainty and the total uncertain volume.

Single Region of Uncertainty. Imagine the segmentation of an image with $K = 3$ independent regions, α, β and γ, as depicted in Fig. 1 (**A0**). Region α is certainly not part of the structure ($p_\alpha = 0$, i.e. background), region β belongs to the structure with probability p_β and region γ is certainly part of the structure ($p_\gamma = 1$). Let their volumes be $s_\alpha = 100$, s_β, $s_\gamma = 1$, respectively, with $\mu = \frac{s_\beta}{s_\gamma} = s_\beta$ the volume ratio of uncertain to certain part of the structure. Assuming a perfect algorithm, the optimal predictions under the empirical risk from Eq. 11 are:

$$\arg\max_{\hat{p}_\alpha, \hat{p}_\beta, \hat{p}_\gamma}[\mathbf{E}[\frac{2(s_\beta\hat{p}_\beta p_\beta + s_\gamma\hat{p}_\gamma)}{s_\alpha\hat{p}_\alpha + s_\beta\hat{p}_\beta + s_\gamma\hat{p}_\gamma + s_\beta p_\beta + s_\gamma}]] \tag{12}$$

It is trivial to show that $\hat{p}_\alpha = 0 = p_\alpha$ and $\hat{p}_\gamma = 1 = p_\gamma$ are solutions for this equation. The behavior of $\hat{p}_\beta$ w.r.t. p_β and μ can be observed qualitatively in Fig. 1 (**A1-A4**). Indeed, only for $p_\beta = \{0,1\}$ the predicted uncertainty $\hat{p}_\beta$ is exact. The location of the local minimum in $\hat{p}_\beta = [0,1]$ switches from 0 to 1 when $p_\beta = 0.5$. Therefore, when p_β decreases or increases from 0.5 (different opacity in **A1-A3**), respectively under- or overestimation will occur (**A4**). The

resulting volumetric bias will be highest when the inherent uncertainty $p_\beta = 0.5$ and decreases towards the points of complete certainty, being always 0 or 1. The effect of the volume ratio μ (colors) is two-fold. With μ increasing, the optimal loss value increases (**A1-A3**) and the volumetric bias increases (**A4**; solid lines). However, the error on the estimated uncertainty is not influenced by μ (**A4**; dashed lines).

Multiple Regions of Uncertainty. In a similar way we can imagine the segmentation of a structure with $K = N + 2$ independent regions, for which we further divided the region β into N equally large independent sub-regions β_n with $n = 0 \ldots N - 1$. Let us further assume they have the same inherent uncertainty $p_{\beta_n} = p_\beta$ and volume ratio $\mu_{\beta_n} = \frac{\mu_\beta}{N}$ (in order to keep the total uncertain volume the same). If we limit the analysis to a qualitative observation of Fig. 1 with $N = 4$ (**B0-B4**) and $N = 16$ (**C0-C4**), we notice three things. First, the uncertainty p_β for which under- or overestimation will happen decreases (**A4**,
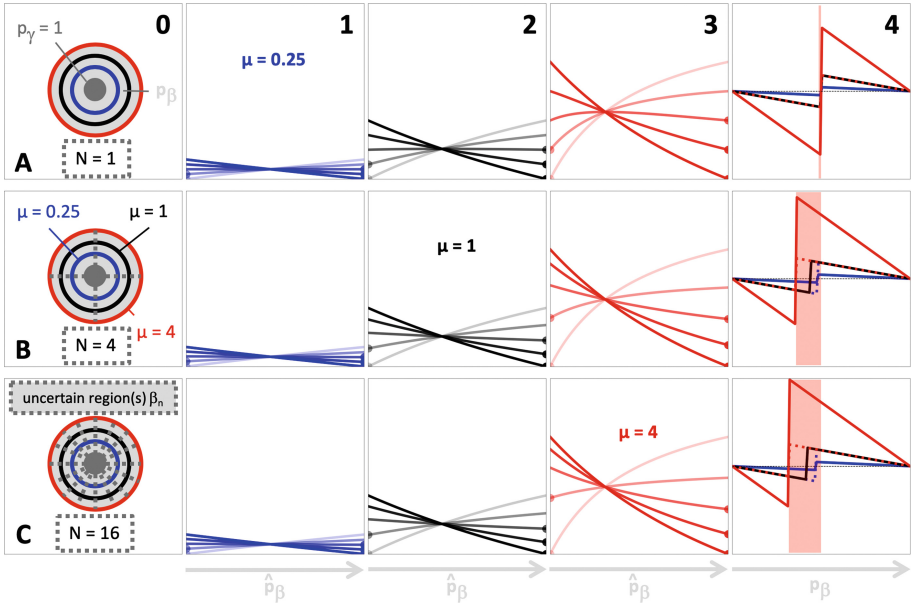
Fig. 1. The effects of optimizing w.r.t. $\mathcal{SD}$ for volume ratios: $\mu = 0.25$ (blue), $\mu = 1$ (black) and $\mu = 4$ (red). ROWS **A-C**: Situations with respectively $N = \{1, 4, 16\}$ independent regions with uncertainty p_β. COLUMN **0**: Schematic representation of the situation. COLUMNS **1-3**: $\mathcal{SD} = [0, 1]$ (y-axis) for $p_\beta = \{0, 0.25, 0.5, 0.75, 1\}$ (respectively with increasing opacity) and $\hat{p} = [0, 1]$ (x-axis). COLUMN **4**: Influence of $p_\beta = [0, 1]$ (x-axis) on volumetric bias (solid lines) or on the error in predicted uncertainty (dashed lines). With the light red area we want to highlight that easier overestimation of the predicted volume occurs due to a higher volume ratio μ or an increasing number of independent regions N. (Color figure online)

B4, **C4**). Second, this effect is proportional with μ and the maximal error on the predicted uncertainty becomes higher (**B0-B4**, **C0-C4**). Third, there is a trend towards easier volumetric overestimation and with the maximal error being more pronounced when the number of regions increases (**A4**, **B4**, **C4**).

3   Empirical Analysis

In this section we will investigate whether the aforementioned characteristics can be observed under real circumstances. In a practical scenario, the joint probability distribution $P(\mathbf{x}, y)$ is unknown and presents itself as a training set. The risk $\mathcal{R}_{\mathcal{L}}$ (Eq. 4) becomes empirical, where the expectation of the loss function becomes the mean of the losses across the training set. Furthermore, the loss $\mathcal{L}$ absorbs the explicit (e.g. weight decay, L2) or implicit (e.g. early stopping, dropout) regularization, which is often present in some aspect of the optimization of CNNs. Finally, the classifier is no longer perfect and additionally to the inherent uncertainty in the task we now have inherent uncertainty introduced by the classifier itself.

To investigate how these factors impact our theoretical findings, we train three models with increasing complexity: LR (logistic regression on the input features), ConvNet (simpler version of the next) and U-Net. We use five-fold cross-validation on the training images from two tasks with relatively low inherent uncertainty (i.e. lower-left third molar segmentation from panoramic dental radiographs (MOLARS) [8], BRATS 2018 [4]) and from two tasks with relatively high inherent uncertainty (i.e. ISLES 2017 [2], ISLES 2018 [3]). Next, we describe the experimental setup, followed by a dissemination of the predicted volume errors $\Delta \mathcal{V}(\hat{Y}, Y) = \mathcal{V}(\hat{Y}) - \mathcal{V}(Y)$ by $\mathcal{CE}$ and $\mathcal{SD}$ trained models.

3.1   Task Description and Training

We (re-)formulate a binary segmentation task for each dataset having one (multi-modal) input, and giving one binary segmentation map as output (for BRATS 2018 we limit the task to whole tumor segmentation). For the 3D public benchmarks we use all of the provided images, resampled to an isotropic voxel-size of 2 mm, as input (for both ISLES challenges we omit perfusion images). In MOLARS (2D dataset from [8]), we first extract a 448×448 ROI around the geometrical center of the lower-left third molar from the panoramic dental radiograph. We further downsample the ROI by a factor of two. The output is the segmentation of the third molar, as provided by the experts. All images are normalized according to the dataset's mean and standard deviation.

For our U-Net model we start from the successful No New-Net implementation during last year's BRATS challenge [12]. We adapt it with three $3 \times 3 (\times 3)$ average pooling layers with corresponding linear up-sampling layers and strip the instance normalization layers. Each level has two $3 \times 3 (\times 3)$ convolutional layers before and after the pooling and up-sampling layer, respectively, with [[10, 20], [20, 10]], [[20, 40], [40, 20]], [[40, 80], [80, 40]] and [40, 20] filters. For the ConvNet

model, we remove the final two levels. The LR model uses the inputs directly for classification, thus performing logistic regression on the input features.

The images are augmented intensively during training and inputs are central image crops of $162 \times 162 \times 108$ (in MOLARS 243×243). We train the models w.r.t. $\mathcal{CE}$ or $\mathcal{SD}$ with ADAM, without any explicit regularization, and with the initial learning rate set at 10^{-3} (for LR model at 1). We lower the learning rate by a factor of five when the validation loss did not improve over the last 75 epochs and stop training with no improvement over the last 150 epochs.

3.2   Results and Discussion

In Table 1 the results are shown for each dataset (i.e. MOLARS, BRATS 2018, ISLES 2017, ISLES 2018), for each model (i.e. LR, ConvNet, U-Net) and for each loss (i.e. $\mathcal{CE}$, $\mathcal{SD}$) after five-fold cross-validation. We performed a pairwise non-parametric significance test (bootstrapping) with a p-value of 0.05 to assess inferiority or superiority between pairs of optimization methods.

Table 1. Empirical results for cross-entropy ($\mathcal{CE}$), soft Dice score ($1-\mathcal{SD}$) and volume error ($\Delta\mathcal{V}$; in 10^2 *pixels* or *ml*) metrics for models optimized w.r.t. $\mathcal{CE}$ and $\mathcal{SD}$ losses. Significant volumetric underestimations in *italic* and overestimations in **bold**.

Dataset ↓	Model →	LR		ConvNet		U-Net	
	Training loss →	$\mathcal{CE}$	$\mathcal{SD}$	$\mathcal{CE}$	$\mathcal{SD}$	$\mathcal{CE}$	$\mathcal{SD}$
	Metric ↓						
MOLARS (2D)	$\mathcal{CE}(\hat{Y},Y)$	0.240	5.534	0.194	1.456	0.024	0.103
	$1-\mathcal{SD}(\hat{Y},Y)$	0.068	0.153	0.150	0.270	0.865	0.931
	$\Delta\mathcal{V}(\hat{Y},Y)$ (10^2 *pixels*)	−0.069	**302.3**	−0.276	**87.09**	0.092	−0.187
BRATS 2018 (3D)	$\mathcal{CE}(\hat{Y},Y)$	0.039	0.173	0.030	0.069	0.012	0.027
	$1-\mathcal{SD}(\hat{Y},Y)$	0.080	0.355	0.196	0.715	0.585	0.820
	$\Delta\mathcal{V}(\hat{Y},Y)$ (*ml*)	−2.841	**276.4**	3.936	**19.93**	*−6.778*	−1.905
ISLES 2017 (3D)	$\mathcal{CE}(\hat{Y},Y)$	0.025	0.155	0.018	0.069	0.014	0.066
	$1-\mathcal{SD}(\hat{Y},Y)$	0.099	0.255	0.114	0.321	0.188	0.340
	$\Delta\mathcal{V}(\hat{Y},Y)$ (*ml*)	15.71	**82.42**	−4.227	**23.83**	−2.875	**13.44**
ISLES 2018 (3D)	$\mathcal{CE}(\hat{Y},Y)$	0.055	0.225	0.044	0.139	0.029	0.128
	$1-\mathcal{SD}(\hat{Y},Y)$	0.136	0.329	0.200	0.449	0.362	0.518
	$\Delta\mathcal{V}(\hat{Y},Y)$ (*ml*)	0.773	**34.03**	−0.374	**12.44**	−0.878	**5.442**

Optimizing the $\mathcal{CE}$ loss reaches significantly higher log-likelihoods under all circumstances, while soft Dice *scores* (i.e. $1-\mathcal{SD}$) are significantly higher for $\mathcal{SD}$ optimized models. Looking at the volume errors $\Delta\mathcal{V}(\hat{Y},Y)$, the expected outcomes are, more or less, confirmed. For the LR and ConvNet models, $\mathcal{CE}$ optimized models are unbiased w.r.t. volume estimation. For these models, $\mathcal{SD}$ optimization leads to significant overestimation due to the remaining uncertainty, partly being introduced by the models themselves.

The transition to the more complex U-Net model brings forward two interesting observations. First, for the two tasks with relatively low inherent uncertainty (i.e. MOLARS, BRATS 2018), the model is able to reduce the uncertainty

to such an extent it can avoid significant bias on the estimated volumes. The significant underestimation for $\mathcal{CE}$ in BRATS 2018 can be due to the optimization difficulties that arise in circumstances with high class-imbalance. Second, although the model now has the ability to extend its view wide enough and propagate the information in a complex manner, the inherent uncertainty that is present in both of the ISLES tasks, brings again forward the discussed bias. In ISLES 2017, having to predict the infarction after treatment straightforwardly introduces uncertainty. In ISLES 2018, the task was to detect the acute lesion, as observed on MR DWI, from CT perfusion-derived parameter maps. It is still unknown to what extent these parameter maps contain the necessary information to predict the lesion.

The $\mathcal{CE}$ optimized U-Net models result in Dice scores (Eq. 1) of 0.924, 0.763, 0.177 and 0.454 for MOLARS, BRATS 2018, ISLES 2017 and ISLES 2018, respectively. The Dice scores obtained with their $\mathcal{SD}$ optimized counterparts are significantly higher, respectively 0.932, 0.826, 0.343 and 0.527. This is in line with recent theory and practice from [7] and justifies $\mathcal{SD}$ optimization when the segmentation quality is measured in terms of Dice score.

4 Conclusion

It is clear that, in cases with high inherent uncertainty, the estimated volumes with soft Dice-optimized models are biased, while cross-entropy-optimized models predict unbiased volume estimates. For tasks with low inherent uncertainty, one can still favor soft Dice optimization due to a higher Dice score.

We want to highlight the importance of choosing an appropriate loss function w.r.t. the goal. In a clinical setting where volume estimates are important and for tasks with high or unknown inherent uncertainty, optimization with cross-entropy can be preferred.

References

1. NEXIS - Next gEneration X-ray Imaging System. https://www.nexis-project.eu
2. Ischemic Stroke Lesion Segmentation (ISLES) challenge (2017). http://www.isles-challenge.org/ISLES2017/
3. Ischemic Stroke Lesion Segmentation (ISLES) challenge (2018). http://www.isles-challenge.org/ISLES2017/
4. Multimodal Brain Tumor Segmentation (BRATS) challenge (2018). https://www.med.upenn.edu/sbia/brats2018.html
5. Bakas, S., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**(March), 1–13 (2017). https://doi.org/10.1038/sdata.2017.117

6. Bakas, S., et al.: Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge (2018). http://arxiv.org/abs/1811.02629

7. Bertels, J., et al.: Optimizing the Dice score and Jaccard index for medical image segmentation: theory and practice. In: Medical Image Computing and Computer-Assisted Intervention (2019)

8. De Tobel, J., Radesh, P., Vandermeulen, D., Thevissen, P.W.: An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. J. Forensic Odonto-Stomatol. **35**(2), 49–60 (2017)

9. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: Deep Learning and Data Labeling for Medical Applications. LNCS, vol. 10008, pp. 179–187. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-46976-8

10. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

11. Goyal, M., et al.: Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet **387**(10029), 1723–1731 (2016). https://doi.org/10.1016/S0140-6736(16)00163-X

12. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21. http://arxiv.org/abs/1809.10483

13. Kamnitsas, K., Ledig, C., Newcombe, V.F.J.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)

14. Menze, B.H., et al.: The multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging (2015). https://doi.org/10.1109/TMI.2014.2377694

15. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision, vol. 4, pp. 1–11 (2016). https://doi.org/10.1109/3DV.2016.79. http://arxiv.org/abs/1606.04797

16. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28

17. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995). https://doi.org/10.1007/978-1-4757-3264-1

18. Winzeck, S., et al.: ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. Front. Neurol. **9**(SEP) (2018). https://doi.org/10.3389/fneur.2018.00679