Difficulty Metrics Study for Curriculum-Based Deep Learning in the Context of Stroke Lesion Segmentation

Juliette Moreau, Laura Mechtouff, David Rousseau, Tae-Hee Cho, Omer Eker, Yves Berthezène, Carole Frindel

▶ **To cite this version:**

Juliette Moreau, Laura Mechtouff, David Rousseau, Tae-Hee Cho, Omer Eker, et al.. Difficulty Metrics Study for Curriculum-Based Deep Learning in the Context of Stroke Lesion Segmentation. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Apr 2023, Cartagena, Colombia. pp.1-5, 10.1109/ISBI53787.2023.10230836 . hal-04210497

HAL Id: hal-04210497
https://hal.science/hal-04210497v1

Submitted on 18 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DIFFICULTY METRICS STUDY FOR CURRICULUM-BASED DEEP LEARNING IN THE CONTEXT OF STROKE LESION SEGMENTATION

Juliette Moreau[1] *Laura Mechtouff*[23] *David Rousseau*[4] *Tae-Hee Cho*[23] *Omer Eker*[12]
Yves Berthezène[12] *Carole Frindel*[1]

[1]CREATIS, CNRS UMR5220, INSERM U1206, Université Lyon 1, INSA-Lyon, 69621 Villeurbanne, France
[2]Stroke department, Hospices Civils de Lyon, 69500 Bron, France
[3] CarMeN, INSERM U1060, INRAe U1397, Université Lyon 1, INSA de Lyon, 69621 Villeurbanne, France
[4]LARIS, UMR IRHS INRAe, Universite d'Angers, 49100 Angers, France

ABSTRACT

Brain imaging plays a central role in the management of stroke patients, where the two main modalities are magnetic resonance imaging and computed tomography from which automatic segmentation of the lesion is done to help physicians. However current methods are not yet satisfying as they do not consider the diversity of patients. Curriculum learning is a method in machine learning that consists in introducing training examples progressively according to their difficulty. The objective of this work is to study difficulty metrics to establish an order within the data for curriculum-based stroke lesion segmentation. Three difficulty metrics are tested, lesion area, image contrast and a metric based on gradient loss, for two types of segmentation architectures and two imaging modalities. Although the gradient loss metric is the most correlated with the performance results, curriculum learning with image contrast gives equally good results with an increase in Dice up to 13%.

***Index Terms*—** Curriculum learning, difficulty metric, segmentation, MRI, CT, stroke.

1. INTRODUCTION

Ischemic stroke is a major cause of acquired disability and death. Reperfusion therapies are the current standard-of-care to promote clinical recovery. Brain imaging plays a critical role in patients management. Computed tomography (CT) and magnetic resonance imaging (MRI) are the two main imaging modalities used to establish the diagnosis and specify the size and location of the lesion. MRI is more expensive, time consuming and less available than CT but offers images with better contrast. Having a robust tool to automatically segment the lesion would save clinicians considerable time in assessing patient prognosis and response to treatment. However, this task remains a methodological challenge, as the existence of the Ischemic Stroke Lesion Segmentation Challenge (ISLES) [1] illustrates every year. The U-Net architecture [2] and its variants are so far the most successful approaches in the state of the art. If the performances in MRI are satisfactory (Dice of 0.82 on the last ISLES edition), on the other hand the results are less convincing in CT and are still poorly represented in the literature due to the lower contrast of the stroke lesion. An explanation of the poor results is the fact that stroke is often considered like one disease but actually there is a great diversity of strokes considering the lesion size, its location, its shape [3] or the delay from symptoms onset to imaging [4]. Curriculum learning (CL) is a method that allows to spread out the examples during the learning process given a difficulty measure. Therefore we searched metrics that could explain the difficulty of the task and study the distribution of patients in term of lesion area, contrast, and gradient loss. To the best of our knowledge it has never been used for stroke lesion segmentation except for one approach proposed to ISLES challenge 2017 [5] but on the basis of synthetic data.

2. RELATED WORK

CL was introduced by analogy with the way humans learns step by step in progressive and organized tasks. Initially, vanilla CL [6] was defined as follows: more difficult examples are added over time based on human-defined criteria. Then variations were developed, such as anti-CL [7], where examples are presented from difficult to easy during the training. Many other methods like self-paced learning (SPL) [8] were created where the difficulty is not fixed and the order is not known before the learning: it is fixed according to the performances of the model. Recently, master-student CL [9] is widely used in which an auxiliary model learns the best parameters for the main network. Hence, the key in CL is to define the difficulty metric of the training examples.

We used vanilla CL model based on difficulty metrics inspired by clinical priors that are lesion area –as we work in 2D to expand the training set– and image contrast. We compared them to a well known difficulty metric in deep learning which is the gradient of the loss as defined in [10].

3. MATERIAL AND METHODS

3.1. Difficulty metrics

To study the heterogeneity of the dataset and understand which slices would be more complicated to segment, we introduced three difficulty metrics. The first is lesion area based on FLAIR MRI masks and calculated for each slice thanks to pixels size referenced in the header of the images.

For the contrast of the lesion, as Gaussian distribution of FLAIR MRI and CT images have been demonstrated [11], we used the Fisher ratio: $F = \frac{(\mu_h - \mu_l)^2}{\sigma_h^2 + \sigma_l^2}$ where μ_h is the mean value of pixels in healthy tissue of hemisphere of the lesion (respectively μ_l, in the lesion), and σ_h^2 (and σ_l^2) the standard deviation associated. It represents the difference between the lesion and the healthy tissue according to their intensity taking into account the homogeneity of intensity in each tissue.

To evaluate the difficulty of slices through gradient descent, we used weights from already trained models (U-Net and Mask R-CNN introduced in 3.2) on our data to initialize new models that are retrained with all the slices. We kept the loss variation induced by each slice at each iteration during this second training with the idea that a large loss decrease means that the slice brings more information and is considered easy [10]. We used models that had already converged because randomly initialized models would have much larger loss differences at the beginning of training and the order of the slices would have a huge impact on the difficulty assessment. The experiment was repeated five times to further moderate the effect of slice order. In addition, this second training was run over 100 epochs to collect a sufficient number of values of loss variation per slice and during the whole training process. For each slice, we averaged the change in loss across epochs and for each fold, then they were normalized and a final score - called gradient difficulty - was calculated per slice.

As seen in Figure 1, the distribution of lesion area is very uneven: a majority of images show a very small lesion and are more difficult to segment. As for lesion area, the Fisher ratio distribution in CT is also monotonic whereas in FLAIR MRI it is Gaussian, which reveals one of the major discrepancies between the two imaging modalities. The distribution is also Gaussian for the gradient difficulty. In both imaging modalities, there is a core group of slices that have a similar value of gradient difficulty, but some slices stand out from the rest. These are not necessarily from the same patients in both modalities but all are from patients with smaller lesions.

3.2. Models

Two different neural networks architectures were studied to observe whether CL would be more effective. The first one is the U-Net architecture [2], a multi-scale convolutional neural network – today the most used in segmentation – used as a reference method. Probability maps with two classes (background and lesion) are produced, thresholded at 0.5 to obtain the final segmentation. The second method performs segmentation in two steps: first object detection with bounding boxes around the regions of interest, and then segmentation within the box. There are two main architectures to perform this type of task: Mask R-CNN [12] and poly-YOLO [13]. The first one was chosen because the segmentation output is more accurate since it is not based on bounding polygons.

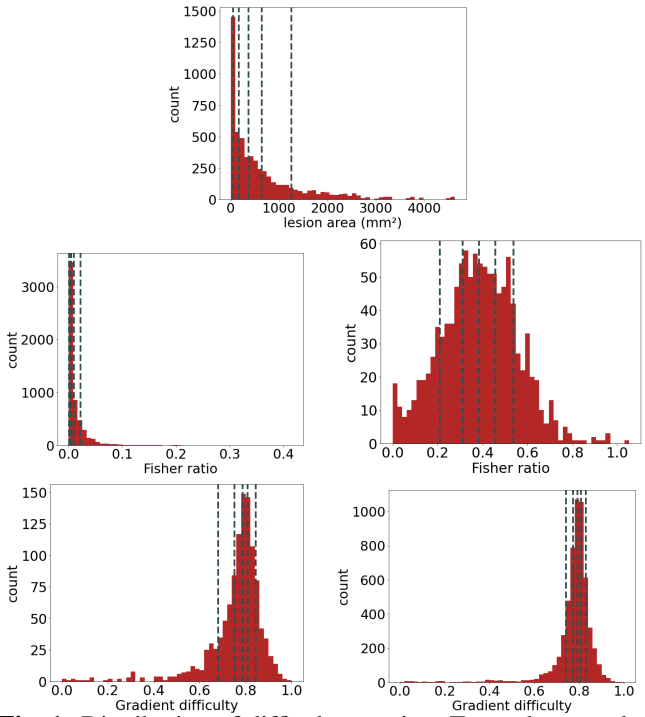

Fig. 1: Distribution of difficulty metrics. Top to bottom: lesion area, Fisher ratio, gradient difficulty. Left: CT images, right: FLAIR MRI. Vertical dotted lines represent quantiles for 6 groups based on the difficulty metrics.

4. EXPERIMENTAL SETUP

4.1. Dataset

We used the single-center HIBISCUS-STROKE cohort [14] for this study. Inclusion criteria were (1) patients with anterior circulation stroke related to proximal intracranial occlusion, (2) diffusion MRI as baseline imaging, (3) patients treated with thrombectomy, (4) CT imaging 24 hours after inclusion, (5) FLAIR MRI 6 days after inclusion from which the ground truth for both CT and MRI is extracted by an expert (THC) with a semi-automatic method [15] as the lesion is more visible and easier to segment manually on it.

Pre-processings were done before training: (1) skull stripping with FSL improved for CT images [16] [17] and HD-BET [18] for MRI, (2) non linear registration on DWI MRI as reference frame with ANTs [19], (3) separation of 3D volumes into 2D slices ; only the slices with lesion according to

ground truth are kept, (4) selective horizontal flipping to place all the lesions in the same hemisphere to establish a prior position of the lesion, (5) resizing to 192x192 pixels without cropping, (6) grayscale normalization. Steps 3 and 4 are based on the assumption that a rough location of the lesion is known from the preliminary clinical examinations. In the end, 108 patients were included which represents 3887 CT images and 753 MRI slices to be segmented.

4.2. Implementation

The original architecture of U-Net [2] with 5 layers was used for the experiments with the sum of cross-entropy and Dice between two classes as loss function. The optimizer is RMSprop [20]. For Mask R-CNN, detectron [21] implementation was used. The loss of object detection is a smooth L1-loss and the one for the segmentation is a sigmoid cross entropy. The optimizer applied is SGD [20] for both steps of learning. The ResNet50 [22] weights were used for object detection and initialized with a model trained with ImageNet [23].

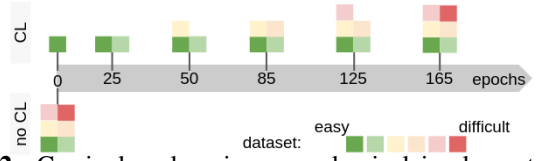

Fig. 2: Curriculum learning vs. classical implementation: more difficult data is added to training through the epochs.

For each experiment (two architectures and two modalities), a 5-fold cross-validation is performed: the dataset is separated into 5 groups of patients, each used once as a validation set. The training lasts 200 epochs with 12 images per batch. When applying CL as represented in figure 2, we kept these parameters but adapted the distribution of images. For each difficulty metric, the training data is separated into 6 groups. For the first 25 epochs, only one-sixth of the slices are given for training. The second group is added to the previous one for the next 25 epochs, and the same process is applied with epoch gaps of 35, 35, 40, and 40, leading to 200 epochs. A non-uniform distribution of epochs was chosen because initially the images are easier to learn.

The evaluation is done thanks to the detection rate which reports the ratio between the number of slices on which a prediction is made via a model on the number of slices with a lesion according to the ground truth, and two other metrics to evaluate the quality of the segmentation which are the Dice score and the Hausdorff distance (HD).

5. RESULTS

5.1. Correlation of performances with difficulty metrics

Prior to CL, we trained the models in a classical manner. The associated results were used to calculate the gradient difficulty metric with the method detailed in 3.1. We then studied the performance on each slice considering its difficulty according to the three chosen metrics of lesion area, contrast and gradient difficulty. Table 1 gathers the R^2 coefficients between the performances (Dice and HD) of each patient and the difficulty metric for the two imaging modalities. We hypothesize that performance increases when a slice is easier according to the difficulty metric (larger lesion, better contrast, or higher loss gradient), which means that Dice is higher and positively correlated with the difficulty metric, whereas HD is smaller and negatively correlated. The closer the R^2 is to 1, respectively -1, the better the correlation and thus the metric could have an impact when using CL.

		area		contrast		gradient difficulty	
		CT	FLAIR	CT	FLAIR	CT	FLAIR
Dice	U-Net	0.42	0.43	-0.01	0.50	0.30	0.69
	R-CNN	0.43	0.40	-0.03	0.54	0.33	0.65
HD	U-Net	0.14	-0.12	-0.04	-0.23	-0.20	-0.25
	R-CNN	0.08	0.01	-0.004	-0.32	-0.24	-0.28

Table 1: R^2 between performances and difficulty metrics.

Regarding lesion area, the Dice score is positively correlated with area with an R^2 of about 0.4 ; segmentation is easier for patients with larger lesions. However, this is not consistent with the fact that HD increases with lesion area for CT images while we expected it to decrease, this phenomenon is less marked for FLAIR MRI. This could be related to the definition of HD, which corresponds to the greater distance between ground truth and segmentation, so for larger lesions the probability of being further from ground truth is higher. While when the contrast increases, the HD decreases and the Dice score improves. For FLAIR MRI, the R^2 is 0.52 for Dice and -0.28 for HD, whereas for CT there is no correlation (R^2 of -0.02 for Dice and HD). The distribution of the Fisher ratio could be one of the reasons: the vast majority of slices correspond to very small values and this dense region drastically reduces the correlation (figure 1); this difference between modalities is shown in figure 4. Finally for the gradient based-difficulty metric (right part of Figure 4), it is more correlated with performance than lesion area or contrast, which can be explained by the fact that the metric is derived from the machine learning model itself.

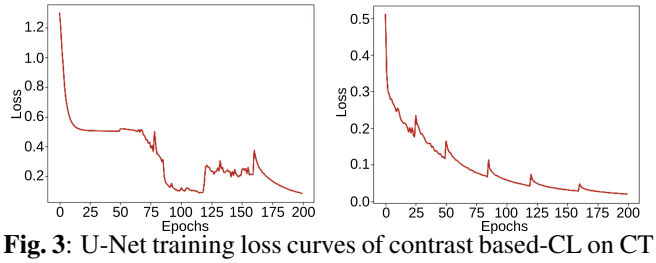

Fig. 3: U-Net training loss curves of contrast based-CL on CT images with two patients separation strategies: left separation according to difficulty steps and right separation in quantiles.

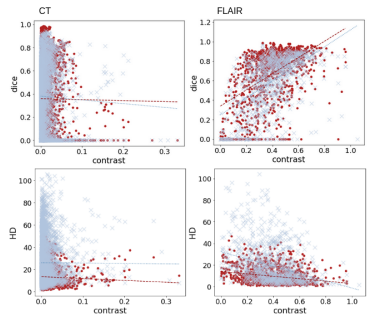

Fig. 4: Correlation between the performances and the contrast.
Red: U-Net, blue: Mask R-CNN.

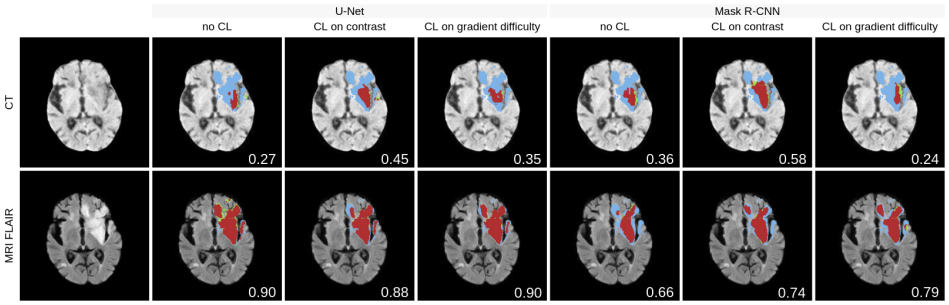

Fig. 5: Segmentation results. Red: true positive, blue: false negative, green: false positive. Number under each slice is the DSC. Segmentation is better for MRI, CL increase the performances.

		U-Net			Mask R-CNN		
		no CL	CL on contrast	CL on gradient difficulty	no CL	CL on contrast	CL on gradient difficulty
CT	detection	80±5	**82±3**	79±4	**75±2**	67±10 **	71±6 **
	dice	0.34±0.04	0.46±0.09 **	**0.47±0.1** **	0.35±0.05	**0.41±0.07** *	0.39±0.07
	HD	30±4	**29±11**	30±6	**26±2**	29±4 **	29±4 **
FLAIR	detection	**97±1**	94±3 **	94±3 **	95±2	97±1 **	**98±1** ***
	dice	0.66±0.04	**0.79±0.02** ***	**0.79±0.02** ***	0.59±0.03	**0.60±0.03**	**0.60±0.03**
	HD	20±2	17±4 *	**16±3** ***	**19±2**	20±1	**19±2**

Table 2: Results of CL with two difficulty metrics. One-tailed unpaired Mann-Whitney statistical test if a CL method is significantly different from classical method (*** if $p < 0.01$, ** if $p < 0.05$, * if $p < 0.08$).

5.2. Curriculum learning

Given the previous results, we applied the CL strategy as explained in 4.2. We decided to divide the patients into 6 quantiles of the same size even if the differences in difficulty metric between the groups could be very unequal rather than selecting patients on the basis of homogeneous steps in difficulty metric. We tested both methods, all results are not presented here, and the second one did not allow a stable convergence as represented in figure 3 for the CT images.

Since the correlation between HD and lesion area was inversely proportional, we decided to eliminate it and focus on the other two difficulty measures. Gradient difficulty seems more promising to have an impact in CL because it correlates well with performance. But we also kept contrast because it was strongly related to FLAIR MRI. The first gain of CL is the training time: there are 73600 iterations without CL for 200 epochs and a batch size of 12 while there are about 47000 with CL (depending on the difficulty metric) because the images are added incrementally. All results are depicted in table 2 and some visualization are given in figure 5.

In all cases, CL allows to have better Dice, the best increase of 0.13 is obtained with U-Net ($p = 0.003$ with FLAIR MRI and $p = 0.024$ with CT images). But for FLAIR MRI this implies a degradation of the detection rate whereas with the Mask R-CNN this rate is increased using CL with FLAIR MRI. This significant improvement, obtained for FLAIR images, is probably explained by the nature of the distribution of difficulty metrics for this modality, where groups of patients can be of similar size and with a homogeneous difference in difficulty. The two selected difficulty metrics (contrast and gradient difficulty) produce similar results. From a computational cost perspective, the contrast-based one is the more interesting as it does not require additional training and is directly related to the image properties.

6. CONCLUSION

As all stroke cases cannot be considered equally, we evaluated the difficulty associated with each slice for brain lesion segmentation according to three difficulty metrics and applied CL to improve the segmentation results. Considering correlation between performance and difficulty, gradient difficulty is the best. When we look at segmentation performances both gradient difficulty-based and contrast-based CL are effective with a promising increase of quality of segmentation. This work also provides a broader view of what a good difficulty metric for CL is and opens perspectives to other clinical applications than stroke.

7. ACKNOWLEDGMENTS

This work was supported by the RHU MARVELOUS (ANR-16-RHUS-0009) of Universite Claude Bernard Lyon-1 (UCBL) and by the RHU BOOSTER (ANR-18-RHUS-0001), within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR).

8. COMPLIANCE WITH ETHICAL STANDARDS

This study followed the principles of the Declaration of Helsinki, and was approved by the local ethics committee (IRB number: 00009118, 2016 September 8th). All subjects or their relatives signed an informed consent form.

9. REFERENCES

[1] O. Maier et al., "Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri," *Medical image analysis*, vol. 35, pp. 250–269, January 2017.

[2] 0. Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, vol. 18, pp. 8234–241.

[3] C. Frindel et al., "Validity of shape as a predictive biomarker of final infarct volume in acute ischemic stroke," *Stroke*, vol. 46, pp. 976–981, March 2015.

[4] J.M. Wardlaw et al., "The impact of delays in computed tomography of the brain on the accuracy of diagnosis and subsequent management in patients with minor stroke," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, pp. 77–81, January 2003.

[5] X. Hu et al., "Strokenet: 3d local refinement network for ischemic stroke lesion segmentation," in *Int. MICCAI Brainlesion Workshop*. MICCAI.

[6] Y. Bengio et al., "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ICML, vol. 26.

[7] S. Braun et al., "A curriculum learning method for improved noise robustness in automatic speech recognition," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, vol. 25, pp. 548–552.

[8] M. Kumar et al., "Self-paced learning for latent variable models," in *Advances in neural information processing systems*. NIPS, 2010, vol. 23.

[9] T.-H. Kim and J. Choi, "Screenernet: Learning self-paced curriculum for deep neural networks," *arXiv preprint arXiv:1801.00904*, June 2018.

[10] G. Hacohen and Da. Weinshall, "On the power of curriculum learning in training deep networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2535–2544.

[11] P. Gravel et al., "A method for modeling noise in medical images," *IEEE Transactions on medical imaging*, vol. 23, pp. 1221–1232, October 2004.

[12] K. He et al., "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*. IEEE, 2017, pp. 2961–2969.

[13] Hurtik P. et al., "Poly-yolo: higher speed, more precise detection and instance segmentation for yolov3," *Neural Computing and Applications*, vol. 34, pp. 8275–8290, February 2022.

[14] N. Debs et al., "Impact of the reperfusion status for predicting the final stroke infarct using deep learning," *NeuroImage: Clinical*, vol. 29, pp. 102548, January 2021.

[15] A. Fedorov et al., "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30(9), pp. 1323–41, November 2012.

[16] M. Jenkinson et al., "Fsl," *NeuroImage*, vol. 62, pp. 782–90, August 2012.

[17] J. Muschelli et al., "Validated automatic brain extraction of head ct images," *NeuroImage*, vol. 114, pp. 379–85, July 2015.

[18] F. Isensee et al., "Automated brain extraction of multisequence mri using artificial neural networks," *Human Brain Mapping*, vol. 40, pp. 4952– 4964, August 2019.

[19] B.B. Avants et al., "Advanced normalization tools (ants)," *Insight*, vol. 2, pp. 1–35, January 2009.

[20] D. Choi et al., "On empirical comparisons of optimizers for deep learning," *CoRR*, vol. abs/1910.05446, October 2019.

[21] R. Girshick et al., "Detectron," https://github.com/facebookresearch/detectron, 2018.

[22] S. Targ et al., "Resnet in resnet: Generalizing residual architectures," March 2016.

[23] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, April 2015.