2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC) | 978-1-6654-8872-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/NSS/MIC44845.2022.10399232

# Curriculum Learning for Improved Tumor Segmentation in PET Imaging

Fereshteh Yousefirizi, Carlos Uribe, and Arman Rahmim, *Senior Member, IEEE*

*Abstract*– **In this paper, we considered the effect of non-uniform sampling of the training data i.e. curriculum learning (CL) on the performance of 3D convolutional neural network for tumor segmentation in PET images. We applied two different curriculums for providing the training data to a convolutional neural network (a 3D U-Net with squeeze and excitation normalization). We applied the easy to hard curriculums by (i) bootstrapping scoring and (ii) self-paced scoring functions. We used augmentation of training data to improve the generalization capability of the trained model and focal loss to take into account the rare samples in any curriculum of training. The learning curves showed that the curriculums based on bootstrapping and self-paced functions speed up the learning while reverse ordering (anti-scoring) makes the training slower and degrades the test performance. The learning by random (uniform) curriculum converges slower. The segmentation results on test data showed that an effective CL via bootstrapping improved segmentation performance and outperformed the trained models obtained via SPS and random curriculums (Dice_bootsrtapping=0.78±0.05 vs. Dice_random=0.72±0.17 vs. Dice_sel-paced=0.63 ± 0.1). In fact, SPS based approach showed a lower mean dice score compared to random curriculum. Anti-Scoring curriculum had the lowest performance in terms of Dice score (0.51± 0.21), that confirmed the effect of curriculum on the learning performance.**

## I. Introduction

SEGMENTAITION of tumors from positron emission tomography (PET) images is mostly needed for radiomics and quantitative PET analysis. Convolutional neural networks (CNNs) have shown good performance for medical image interpretation [1], [2]. However, it usually involves dense training while data unbalance often happens. The propagated information within the network during the early stages of training a CNN can affect the training efficiency.

Motivated by the way humans learn, recent works have shown the importance of curriculum for learning by gradually increasing the difficulty of the learning task [3]; mostly on standard vision datasets [4] and in some medical images such as MRI [5], [6]. Imposing a curriculum to manage the learning process by adjusting the order and pace of training data can speed up learning, improving the final performance and generalization compared to the same model trained without curriculum. Furthermore, the issue of dense training and class imbalance can be handled effectively and efficiently by curriculum learning (CL) [7].

Curriculum learning (CL) needs a method for assessing the level of difficulty of the training data. However, this can be challenging because the assessment measure is not obvious, and an ideal difficulty ranking function is not always available. A curriculum can be designed by measuring the complexity of training data using its distribution density in feature space, and rank the complexity in an unsupervised [8] or supervised manner [4].

Hu et al. [5] suggested a CL based on the assumption that the level of complexity increases gradually by adding the augmented data to the training set and applying focal loss that encourages the segmentation model to learn from hard samples (rare samples with more complexity) of MRIs. However, this curriculum cannot guarantee the performance improvement and their definition of complexity is subjective.

A curriculum can be defined by prioritizing the higher loss samples—which are more likely to come from minority classes—priority by reweighting training samples, which reduces the bias caused by class imbalance. Additionally, CL may favor cases with lower losses since they are more likely to have clean data, which lessens the bias caused by label noise [7].

In this regards, two scoring approaches for CL were introduced previously [4]: i) bootstrapping scoring function and ii) self-paced scoring function. Our goal is to assess and ascertain the importance of learning strategy and to design a learning curriculum by bootstrapping and self-paced functions for automatic segmentation of gross tumor volume of head and neck oropharyngeal primary tumors in PET images.

## II. Material and Methods

### A. Data and augmentation approaches

PET images of 224 patients that received a scan with [$^{18}$F]FDG for head and neck cancer at five institutions were used. PET scans and corresponding ground truths for the primary gross tumor volume were provided by HECKTOR

This work was supported by the Canadian Institutes of Health Research (CIHR) Project Grant PJT-173231.

Fereshteh Yousefirizi is with the Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, Canada (e-mail: frizi@bccrc.ca ). Carlos Uribe is with BC Cancer, Vancouver, Canada (e-mail: curibe@bccrc.ca).

Arman Rahmim is with the Department of Radiology, and Department of Physics & Astronomy, University of British Columbia, Vancouver, Canada. He is also with the Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, Canada (e-mail: arman.rahmim@ubc.ca)

challenge 2021 [9]. The following augmentation techniques were applied to increase the complexity of the training data: i) spatial and intensity transformation (i.e. rotation in random direction (< 25 degree), ii) scaling with a random factor (0.8 and 1.2), iii) elastic deformations, iv) Gamma corrections with γ sampled from the uniform distribution (0.8 and 1.2)), and v) mixup [10].

The mixup augmentation approach involves two input images and their corresponding targets that are proportionally interpolated (dropout should be used to overcome overfitting). Mixup augmentation extends the training distribution by linear interpolations of training samples that leads to the linear interpolations of their corresponding targets (we used one-hot target vectors). We also applied affine registration before mixup augmentation to generate meaningful augmented training data (e.g. avoid an augmented case with two heads).

### B. Segmentation Model

For training, we utilized a 3D U-Net model (Fig. 1) with residual layers, supplemented with squeeze and excitation (SE) normalization and learnable non-linear downsampling and upsampling branches. The core component of the 3D U-Net is made up of the standard convolutional blocks, including a 3×3×3 convolution, a batch normalization layer, and a ReLU activation unit.

Fig. 1 shows how we used the residual blocks in addition to a concurrent spatial and channel squeeze and excitation module SE normalization (Fig. 1 green blocks). The feature maps are given weight by the squeeze & excitation modules so that the network can adaptively focus its attention.

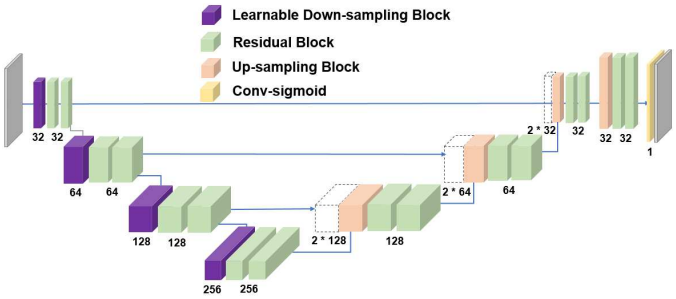


Fig. 1. 3D U-Net with squeeze and excitation modules .The upsampling blocks in the decoder of the network were implemented by a 3×3×3 transposed convolution instead orange blocks.

There are three main categories of loss functions that have been used for medical image segmentation: distribution-based losses (e.g. cross entropy, Focal loss [11]), region-based losses (e.g. Dice), boundary-based loss (e.g. Mumford-Shah [12]). We showed that for tumor segmentation from PET scans the combination of loss functions from these categories improved the performance [13], [14]. In this work, we used a loss function composed on Focal, Dice and Mumford-Shah losses.

### C. Training Process

The model was trained Adam optimizer (with $\beta 1$=0.9 and $\beta 2$=0.99 for the exponential decay rates for moment estimates) on two NVIDIA Tesla V100 GPUs 16 GB with a batch size of 2. The cosine-annealing schedule was applied to reduce the learning rate from $10^{-3}$ to $10^{-6}$ within every 25 epochs and performing the adjustment at each epoch.

### D. Curriculum learning

Both of the bootstrapping and self-paced functions are defined based on the loss of training data with respect to the desired output. In bootstrapping the desired output is defined with respect to target while in self-paced the scoring function is determined based on the loss of each training data with respect to model prediction [7]. The training scheduler then selects a batch of training data from the comparatively simple examples (based on bootstrapping or self-paced) and delivers it to the model trainer for training at each training epoch. Fig. 2 shows a schematic comparison between random (uniform) training scheme and curriculum learning.

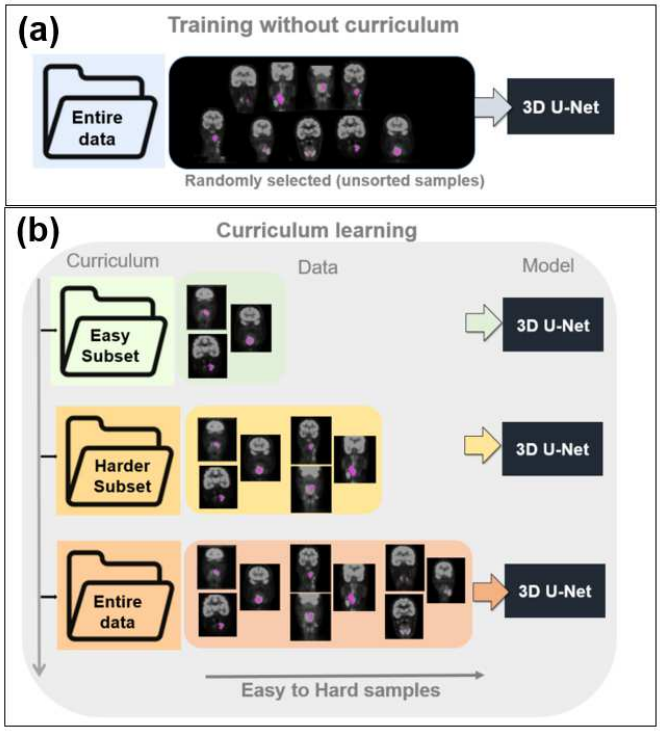


Fig. 2. Schematic of the two training strategies: (a) training with random(uniform) sampling of the training data (b) simple illustration of curriculum learning. CL gradually adds more difficult instances to the subset of training data before training the model using the entire training dataset.

In our learning curriculums, we ranked the difficulty level of the training data, where higher priorities are given to training examples that have lower cost by using (i) bootstrapping and (ii) self-paced functions. For better comparison, we also applied (iii) random ordering and (iv) anti-scoring (using the reverse order of what bootstrapping suggested).

## III. RESULTS

Fig. 1 shows the training curves (dice metric as a function of epoch number) obtained after training the same network without curriculum (i.e. the data are randomly sampled (scored)), as well as curriculums with bootstrapping and self-paced functions, and an anti-scoring curriculum (shown in purple in Fig. 3) converged faster during training. Fig. 3 shows that the curriculum based on bootstrapping and self-paced functions makes the learning start faster, and anti-scoring converges to the worst solution that has been shown in more details in Fig. 4 by comparing the mean dice metric in each curriculum over the last epoch (after five repetitions over different splits of data).

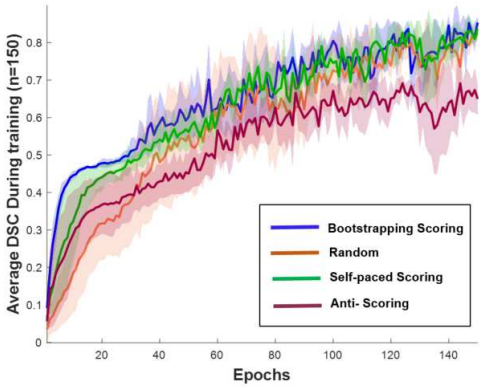


Fig.3.Learning curves of curriculum learning with bootstrapping scoring function method (in violet) converges faster to a higher dice metric during training.

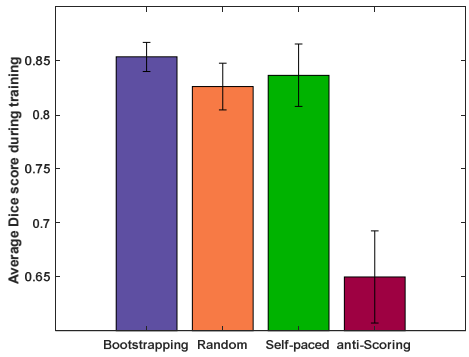


Fig. 4. The mean dice in each curriculum in the last epoch after 5 repetitions over different splits of data. Error bars indicate the standard deviation (STD) of the dice score during training in the last epoch.

Segmentation performance in terms of mean dice score on the test data (chosen randomly from different centers) is summarized in Fig. 4, showing improved performance for the bootstrapping method. The lower mean and higher standard deviation (STD) of the anti-scoring curriculum in the last epoch emphasizes the effect of curriculum on the convergence of segmentation model. Fig. 5 shows that the bootstrapping CL improved the segmentation performance on test data confirming that the use of an effective curriculum for training can be impactful on training process even in the absence of large training datasets.

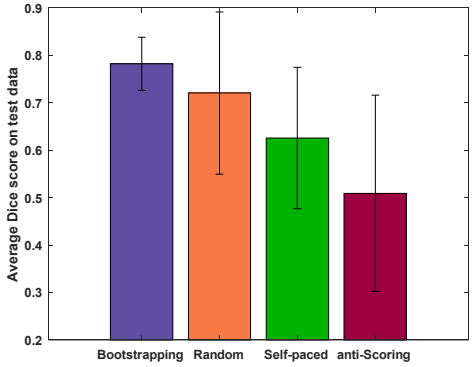


Fig. 5. The mean dice scores for all n=50 test PET images. Error bars indicate the standard deviation (STD) of dice scores.

A one-way repeated measures ANOVA was conducted on test results (n=50) to examine the effect of the four curriculums on segmentation performance. Results showed that the curriculums led to statistically significant differences in dice scores ($p<0.001$). Kolmogorov-Smirnov test ($\alpha = 0.05$) indicated that the different performances of CL approaches were all statistically different from each other (p-value<0.05), except that self-paced and anti-scoring were not statistically different from each other. (Fig. 3) bootstrapping was statistically significantly better than all other curriculum approaches (p-value<0.05).

## IV. DISCUSSION AND CONCLUSION

In most of the CL strategies, training starts on easier data to reach a faster learning, assuming that most of the training time wastes on hard samples. In the other words, the distribution of training and test data have respectively high-confidence and low-confidence in easy and hard samples. To reduce the negative effects of low-confidence samples, such as poor performance or slow convergence, curriculum should be designed for learning from easy examples. This simulates learning from high-confidence common region. An appropriate curriculum can also direct the training method for adaptation to the goal distribution when the target distribution is different from the training distribution as we have PET scans from different institution in the external test set.

Some other difficulty measures can be defined based on signal intensity [15] and annotator agreement [16], that are designed for medical imaging data. Signal intensity can be regarded as a measurement for the informativeness of data features. For example in the task of thoracic disease diagnosis, more severe symptoms provide more information and are easier to recognize. Moreover, annotator agreement [16] is proposed to measure the difficulty of an image determined by annotator agreement. In the other words, we can use the levels of agreement for each image as a measure for the complexity of that image because medical image databases can be annotated by a number of physicians.

We aimed at applying CL to train the segmentation network efficiently when training data dominated by easy samples that is the case with tumor segmentation in PET images. We considered the effect of learning curriculums on the

discriminative and generalization power of the trained segmentation model to assess and ascertain the importance of learning strategy in addition to the architecture design to help CNNs to learn better representations from imbalanced data. We considered two CL approaches under the assumption that increasing the difficulty of images presented to the network (i.e. from easy to hard training samples) improves the training performance. We compared the effect of bootstrapping and self-paced curriculums on the convergence and test performance to uniform (random) and anti-scoring ordering.

Bootstrapping and self-paced approaches have two main benefits compared to predefined CL are as follows: 1) Both of them are semi-automated with a loss-based automatic difficulty measure and a dynamic curriculum, that makes them more adaptable and flexible for different tasks and data distributions. 2) They can be used widely as a plug-in tool since it incorporates the curriculum design into the original deep learning objectives.

Despite the faster convergence of self-paced curriculum, the test performance of random curriculum (mean dice score) is better, surprisingly (Fig. 5). However, the STD of random curriculum results (STD=0.17) are higher than self-paced curriculum (STD=0.1) (p-value<0.05) emphasizing the importance of the complexity measure and scoring definition. The poorest segmentation performance resulted from the anti-scoring curriculum. Overall, best performance was obtained with the segmentation model trained by bootsrapping curriculum. CL can improve the performance of 3D U-Net for tumor segmentation in PET images provided with the appropriate scoring function. Our results revealed that scoring based on target (bootstrapping) performed better compared to ordering the data based on model prediction (self-paced).

Moreover, CL can help training on heterogeneous and noisy training data in which the noisy training samples are not considered in the initial stages of training. We will extend this analysis by considering different complexity measure based on the inter-annotator agreement and the curriculum that is learned from the unsupervised clustering of data. In our upcoming studies, the impact of pacing approaches on learning effectiveness will also be taken into account.

## ACKNOWLEDGMENT

This research was supported by the Canadian Institutes of Health Research (CIHR) Project Grant PJT-173231, in part through computational resources and services provided by Microsoft for Health.

## REFERENCES

[1] F. Yousefirizi, A. K. Jha, J. Brosch-Lenz, B. Saboury, and A. Rahmim, "Toward High-Throughput Artificial Intelligence-Based Segmentation in Oncological PET Imaging," *PET Clin.*, vol. 16, no. 4, pp. 577–596, Oct. 2021.

[2] F. Yousefirizi, Pierre Decazes, A. Amyar, S. Ruan, B. Saboury, and A. Rahmim, "AI-Based Detection, Classification and Prediction/Prognosis in Medical Imaging:: Towards Radiophenomics," *PET Clin.*, vol. 17, no. 1, pp. 183–212, Jan. 2022.

[3] S. Sinha, A. Garg, and H. Larochelle, "Curriculum By Smoothing," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21653–21664, 2020.

[4] G. Hacohen and D. Weinshall, "On The Power of Curriculum Learning in Training Deep Networks," in *Proceedings of the 36th International Conference on Machine Learning*, 09--15 Jun 2019, vol. 97, pp. 2535–2544.

[5] X. Hu *et al.*, "Brain SegNet: 3D local refinement network for brain lesion segmentation," *BMC Med. Imaging*, vol. 20, no. 1, p. 17, Feb. 2020.

[6] Z. Liu *et al.*, "Style Curriculum Learning for Robust Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 2021, pp. 451–460.

[7] X. Wang, Y. Chen, and W. Zhu, "A Survey on Curriculum Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4555–4576, Sep. 2022.

[8] S. Guo *et al.*, "Curriculumnet: Weakly supervised learning from large-scale web images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.

[9] V. Andrearczyk *et al.*, "Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images," in *Head and Neck Tumor Segmentation and Outcome Prediction*, 2022, pp. 1–37.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv [cs.LG]*, 25-Oct-2017.

[11] Lin, Goyal, Girshick, and He, "Focal loss for dense object detection," *Proc. Estonian Acad. Sci. Biol. Ecol.*, 2017.

[12] B. Kim and J. C. Ye, "Mumford–Shah Loss Functional for Image Segmentation With Deep Learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1856–1866, 2020.

[13] F. Yousefirizi *et al.*, "Segmentation and Risk Score Prediction of Head and Neck Cancers in PET/CT Volumes with 3D U-Net and Cox Proportional Hazard Neural Networks," *arXiv [physics.med-ph]*, 16-Feb-2022.

[14] F. Yousefirizi *et al.*, "Convolutional neural network with a hybrid loss function for fully automated segmentation of lymphoma lesions in FDG PET images," in *Medical Imaging 2022: Image Processing*, 2022, vol. 12032, pp. 214–220.

[15] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs," in *Machine Learning in Medical Imaging*, 2018, pp. 249–258.

[16] J. Wei *et al.*, "Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2021, pp. 2473–2483.