Quality-based Regularization for Iterative Deep Image Segmentation

José Rebelo[1], Kelwin Fernandes[2], and Jaime S. Cardoso[3]

Abstract— Traditional image segmentation algorithms operate by iteratively working over an image, as if refining a segmentation until a stopping criterion is met. Deep learning has replaced traditional approaches, achieving state-of-the-art performance in many problems, one of them being image segmentation. However, the concept of segmentation refinement is not present anymore, since usually the images are segmented in a single step. This work focuses on the refinement of image segmentations using deep convolutional neural networks, with the addition of a quality prediction output. The output from a state-of-the-art base segmenter is refined, simultaneously improving it and trying to predict its quality. We show that the quality concept can be used as a regularizer while training a network for direct segmentation refinement.

I. Introduction

Image segmentation consists of partitioning an image in multiple parts, which should have some semantic meaning, i.e. belong to one of multiple classes. Traditional image segmentation algorithms, like region-growing methods, usually operate by iteratively working over preliminary candidate segmentations, optimizing some sort of function or performing some operation until a stopping criterion is reached [1].

Deep Convolutional Neural Networks (CNNs) have shown large success in machine perception tasks, achieving state-of-the-art performance. For image segmentation, CNNs usually consist of encoder-decoder models, which receive an image as input and output the final segmentation [2]. This contrasts with the traditional iterative segmentation methods, which start from scratch or from a coarse segmentation, and iteratively progress towards a finer result.

In this work, we propose the application of iterative refinement CNNs to the task of image segmentation, with the added concept of segmentation quality as a regularizer in the training process. Instead of segmenting an image in a single step, the output from a state-of-the-art network is now used as input for a segmentation refinement deep neural network, as illustrated in Figure 1, which tries to iteratively refine the segmentation while also predicting the quality of the input segmentation. Predicting the quality along with the refined masks acts as a regularizer which promotes the learning of features related to the degree of correction needed to improve the input segmentation.

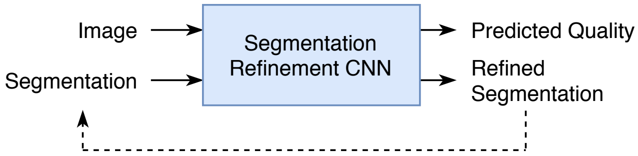

Fig. 1. Proposed segmentation flow

II. Literature Review

Traditional image segmentation methods (those which do not use neural networks) usually involve an iterative process for segmentation refinement. One example consists of the region-growing methods [3]. The segmentation starts from a group of seed pixels, which are iteratively increased by appending to each region neighboring pixels that belong to it, according to a set of defined rules. The region growing stops when no more pixels can be added, according to some stopping criterion. In active contours [4] an image is segmented along the edges, by placing a spline referred to as snake on the image. An energy function is defined, consisting of the sum of the internal and external energies that affect the snake. The internal energy is affected by the deformations made to the snake, while the external energy consists of a combination of the forces caused by the gradients present in the image. Given an initial position for the snake, the energy function is iteratively optimized.

A. Deep learning

Conventional techniques have lately been replaced by deep learning architectures. The most used technique for image segmentation with deep neural networks consists of encoder-decoder architectures. They operate in two phases: first, the input image is encoded into a smaller latent representation, containing some semantic global meaning. That representation is then decoded into the final segmentation. The first example of such an implementation was the SegNet [5]. One evolution of the SegNet is the U-Net [6], which connects each encoding

*This work is co-financed by the ERDF - European Regional Development Fund through the Norte Portugal Regional Operational Programme (NORTE 2020), and the LISBOA2020 under the PORTUGAL 2020 Partnership Agreement and through the Portuguese National Innovation Agency (ANI) as a part of project BCCT.plan: NORTE-01-0247-FEDER-017688.

[1] J. Rebelo is with INESC TEC and University of Porto, Portugal
[2] K. Fernandes is with INESC TEC and NILG.AI, Portugal
[3] J. Cardoso is with INESC TEC and University of Porto, Portugal. For correspondence, use jaime.cardoso@inesctec.pt

level directly to the corresponding decoding level, propagating high resolution features to the decoding process and facilitating the propagation of gradients.

Regularization is one critical strategy to avoid overfitting and improve the generalization capacity of a network [7]. Different regularization methods can be used, either by affecting the weights directly with L1 or L2 regularization, disabling parts of the network to reduce the co-adaptations between the units (e.g. Dropout [8], DropConnect [9], Stochastic Depth [10]), or indirectly, through data augmentation [11]. Data augmentation introduces artificial modifications in the existing training samples, such as rotations, crops, flips and deformations. This improves the generalization performance of a network, and it has been shown that it alone can achieve the same or higher performance, when compared to models trained with other regularization techniques [12].

B. Iterative Segmentation Refinement

There is already some existing work on applying deep learning to the problem of iterative segmentation refinement.

Kim et al. [13] use an encoder-decoder network similar to the U-Net, augmented with an extra input for an already existing segmentation. The existing segmentation is subjected to convolutional layers, before concatenating the obtained feature maps with the ones from the image. A new objective function based on the Dice coefficient is also proposed, which captures the improvement in Dice coefficient between iterations.

Lessmann et al. [14] use CNNs to iteratively segment images of vertebrae. By processing the image in patches from top to bottom, the network retains information about the already segmented vertebrae and uses it to find and segment the next not yet segmented vertebra.

Segmentation methods usually work directly on obtaining output segmentation. Fernandes et al. [15] present a network that infers the quality of a segmentation given an image and segmentation pair. This allows for data augmentation through the unsupervised generation of synthetic segmentations for an image, given that the segmentation quality for the objective function can be easily determined, having the ground-truth before being augmented. With the trained model, it is then possible to iteratively refine a segmentation through backpropagation on the input segmentation, towards a local maximum for the quality.

III. Quality-based Regularization for Image Segmentation

In this section we present our proposal for the direct refinement network. The main idea portrayed in Figure 2 consists of a traditional encoder-decoder architecture, with an extra input for an existing segmentation mask and an output for a predicted quality from the input's image/mask pair, obtained from the latent dimension. The initial segmentation mask can be obtained from any other model or algorithm, and it will be iteratively refined, progressing towards a finer segmentation by then using the networks own output as a new input for further improvement.

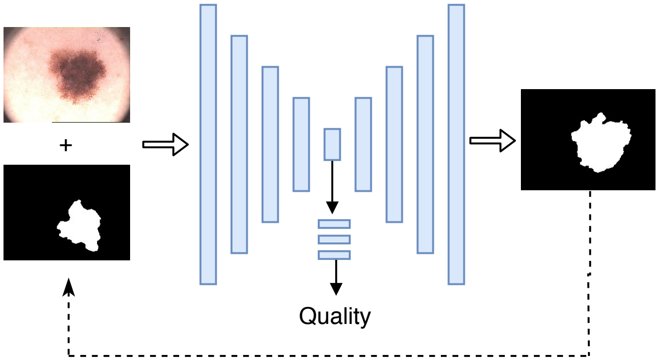

Fig. 2. Proposed model and segmentation flow

The U-Net was used as the base architecture for the direct refinement network. A base segmentation mask, provided by another U-Net trained on the corresponding dataset was concatenated to the input image, as an additional channel.

The quality output extension was added to the U-Net's latent dimension, taking advantage of the semantic information extracted by the encoder before the upsampling/decoding process, as depicted in Figure 3. Global Average Pooling was used to make it independent of input image dimensions, followed by two dense layers. The Dice's coefficient was used as the quality metric. Predicting the quality of the input image-mask pair acts as a regularizer of the segmentation task. Namely, we promote the learning of features related to the errors associated to the current segmentation. Both tasks, image segmentation and quality assessment, are trained in a multitask fashion. Thus, we aim to minimize the Mean Squared Error (MSE) of the quality prediction and to minimize the Dice loss of the output mask (i.e., the complement of the Dice's coefficient). In the loss function, the importance of these two terms is set through a non-negative λ parameter learnt by cross-validation.

IV. Experimental Assessment

All the images are directly used in RGB format, with all the color components normalized to the $[0, 1]$ range. There is minimal preprocessing applied, being just resized to 128×128. For the masks, a binary setting is considered, where pixel values of 0 indicate background and pixel values of 1 indicate foreground (the subject of interest being segmented on each dataset). Training is done for up to a maximum of 500 epochs or 50 epochs without improvement on the validation set, in order to avoid overfitting. Multiple data augmentation techniques

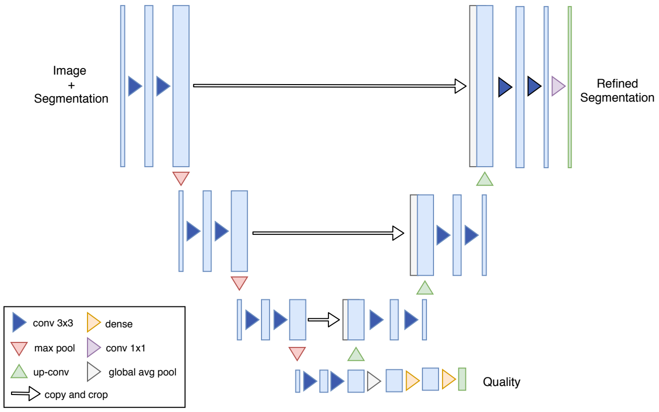

Fig. 3. U-Net for segmentation refinement with quality output extension

were used to augment the input mask, including: elastic deformations, morphological operators (erosion and dilation), random noise, rotations, flips and shifts. We follow the augmentation strategy proposed by Fernandes et al. [15].

The hyperparameter configuration for the models is the one described in Table I, determined to have the best performance using cross-validation. Adam [16] was used as the algorithm for gradient optimization, with a learning rate of $1e^{\;4}$.

TABLE I
Model hyperparameters

Hyperparameter	Value
# Convolution Levels	4
# Consecutive Convolutions	2
# Convolution Filters	32
Convolution Filter Size	3
L2 regularization	0.001
Convolution Activation	ReLU

For the training and evaluation of the proposed solution, we used 6 biomedical datasets that covers the segmentation of skin lesions (PH2 [17] and ISBI 2017 [18]), teeth (Teeth-UCV [19]), breast (Breast-Aesthetics [20]), and cervix (Cervix-HUC [21] and Cervix-MobileODT [22]). We used a traditional 60-20-20 partition for training, validation and test.

A. Results

The performance results for the Refinement U-Net are presented in Table II. All the dice coefficient values in the results have been multiplied by 100, being in the form of a percentage, for easier readability. They were obtained by applying one and two refinement steps to the output of the base U-Net. We can see that the first refinement step provides a big quality increase for all datasets, while a second iteration step starts to perform worse than the previous one for the network trained with no quality output.

The network was then trained with the quality output, using the Dice's coefficient as the prediction quality metric. We can see that the network performs better than the one trained with no quality output in all datasets, and manages to achieve a quality improvement even for a second iteration step. The quality output extension acted as a regularizer, which improved the network's generalization when it was forced to learn the quality alongside the refinement of the segmentation.

Another experiment consisted of introducing already refined outputs as extra training examples, effectively training the network to refine its own output, as an extra data augmentation transformation. This was done up to 2 times, for both training a network from scratch and starting from the existing weights of one trained normally.

The results are shown in Table III. Since it is now possible to refine a segmentation more than twice without the quality starting to decrease (up to 12 times on some datasets), the number of iterations was determined on the validation set, and then using that as stopping criterion for the refinement of the test set, aligned with the work from Fernandes et al. [15]. The best possible refinement is also shown, determined by refining the segmentation until the quality declines when compared with the ground-truth (although unrealistic, since when refining new examples the ground-truth is not known to be used as a stopping criterion, but it serves as a good indicator of the theoretical maximum performance).

The results show that the network that used the pre-trained weights outperformed most of the previous results, without overfitting, especially when using the number of iterations determined on the validation set as stopping criterion. However, it is still evident that a better stopping criterion would be beneficial, since the network falls short of the best possible result in most cases, stopping either too early or too late.

V. Conclusion

This paper proposes the usage of the concept of image quality as a regularizer in the training of deep neural networks for iterative deep image segmentation, as an extension for existing encoder-decoder architectures.

The proposed architecture is validated on several image segmentation datasets, achieving better results when compared to the same setting, without the image quality extension.

As future work, alternative positions and architectures of the quality output extension could be explored, as well as different encoder-decoder architectures apart from the U-Net.

References

[1] P. Alves, J. S. Cardoso, and M. do Bom-Sucesso, "The challenges of applying deep learning for hemangioma lesion segmentation," in Proceedings of the 7th European Workshop on Visual Information Processing (EUVIP), 2018.

TABLE II
Refinement U-Net Performance (Dice coefficient), with and without quality output for 1 (1x) and 2 (2x) refinement iterations. Best result for each dataset highlighted in bold.

Dataset	0x	No Quality		Quality	
		1x	2x	1x	2x
PH2 [17]	83.21	89.10	89.14	90.40	**90.41**
ISBI 2017 [18]	71.30	80.53	79.89	81.13	**81.18**
Teeth-UCV [19]	80.50	83.21	83.62	83.84	**84.83**
Breast-Aesthetics [20]	92.81	93.01	92.77	**94.11**	93.98
Cervix-HUC [21]	76.85	79.10	78.46	79.60	**79.65**
Cervix-MobileODT [22]	87.17	87.98	87.90	88.26	**88.37**

TABLE III
Refinement U-Net Performance, trained with its own output. Scratch corresponds to the network trained from scratch, and Warm to the network trained from the pre-trained weights of another network, trained for just one refinement step. We also show the theoretical best possible result. In parenthesis we show the number of iterations. Best result for each dataset highlighted in bold, not taking into account the best theoretical result.

Dataset	Scratch			Warm	
	2x	Validation	Best	Validation	Best
PH2 [17]	90.41	**90.84** (3)	91.23 (5)	**90.84** (9)	90.88 (11)
ISBI 2017 [18]	81.18	**82.46** (3)	82.68 (1)	80.59 (3)	81.23 (1)
Teeth-UCV [19]	84.83	83.66 (3)	85.30 (7)	**84.87** (3)	86.61 (7)
Breast-Aesthetics [20]	93.98	93.18 (2)	93.35 (2)	**94.17** (3)	94.24 (1)
Cervix-HUC [21]	**79.65**	75.28 (1)	75.29 (1)	78.89 (2)	79.17 (1)
Cervix-MobileODT [22]	88.37	87.08 (2)	87.08 (2)	**88.22** (2)	88.23 (2)

[2] K. Fernandes and J. S. Cardoso, “Ordinal image segmentation using deep neural networks,” in Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2018.

[3] D. D. Patil, S. G. Deore, and S. Bhusawal, “Medical Image Segmentation: A Review,” Ijcsmc, vol. 2, no. 1, pp. 22–27, 2013.

[4] M. Kass, a. Witkin, D. Tetzopoulos, and D. Terzopoulos, “Active contour models,” International Journal of Computer Vision, vol. 1, no. 4, pp. 321–331, 1988.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, dec 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, 2015, pp. 234–241.

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org.

[8] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” CORR, pp. 1–18, 2012.

[9] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, “Regularization of neural networks using dropconnect,” in Proceedings of the 30th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 1058–1066. [Online]. Available: http://proceedings.mlr.press/v28/wan13.html

[10] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, “Deep Networks with Stochastic Depth,” CORR, 2016.

[11] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” CoRR, vol. abs/1712.04621, 2017. [Online]. Available: http://arxiv.org/abs/1712.04621

[12] A. Hernández-García and P. König, “Data Augmentation Instead of Explicit Regularization,” ICLR 2018 Conference, pp. 1–12, 2018.

[13] J. U. Kim, H. G. Kim, and Y. M. Ro, “Iterative deep convolutional encoder-decoder network for medical image segmentation,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 685–688, 2017.

[14] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum, “Iterative fully convolutional neural networks for automatic vertebra segmentation,” CORR, no. Midl, pp. 1–10, 2018.

[15] K. Fernandes, R. Cruz, and J. S. Cardoso, “Deep Image Segmentation by Quality Inference,” Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2018.

[16] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” CORR, pp. 1–15, dec 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[17] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, “PH2- A dermoscopic image database for research and benchmarking,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 5437–5440, 2013.

[18] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 ISBI, hosted by the international skin imaging collaboration (ISIC),” in IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), April 2018, pp. 168–172.

[19] K. Fernandez and C. Chang, “Teeth/palate and interdental segmentation using artificial neural networks,” in Artificial Neural Networks in Pattern Recognition, N. Mana, F. Schwenker, and E. Trentin, Eds., 2012, pp. 175–185.

[20] J. S. Cardoso and M. J. Cardoso, “Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment,” Artificial Intelligence in Medicine, vol. 40, no. 2, pp. 115 – 126, 2007.

[21] K. Fernandes, J. S. Cardoso, and J. Fernandes, “Transfer learning with partial observability applied to cervical cancer screening,” in Proceedings of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), 2017.

[22] K. Inc, “Intel & MobileODT Cervical Cancer Screening,” 2017. [Online]. Available: https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening