U-Net based convolutional neural network for skeleton extraction

Oleg Panichev
Ciklum
Ukraine, Kyiv, Amosova str., 12
pole@ciklum.com

Alona Voloshyna
Ciklum
Ukraine, Kyiv, Amosova str., 12
avol@ciklum.com

Abstract

Skeletonization is a process aimed to extract a line-like object shape representation, skeleton, which is of great interest for optical character recognition, shape-based object matching, recognition, biomedical image analysis, etc.. Existing methods for skeleton extraction are typically based on topological, morphological or distance transform and are known to be sensitive to the noise on the boundary and require post-processing procedure for redundant branches pruning. In this work, we introduce U-net based approach for direct skeleton extraction of the object within Pixel SkelNetOn - CVPR 2019 challenge, inspired by CNNs success in skeleton extraction from real images task. The main idea of our approach is to consistently edit a skeleton mask by feature propagation through different scale layers. It opposes final skeleton generation from different scale object shape representations as occurs in approaches with deep supervision for skeleton extraction from the real image.

Our U-net based model showed 0.75 F1-score on the validation set and the ensemble of eight identical models, trained on different data subsets, got 0.7846 F1-score on the test data.

1. Introduction

Skeletonization is a process aimed to extract a line-like object shape representation, skeleton, allowing the reconstruction of original object shape. [18] gives such skeleton definition: "The skeleton S is a geometric graph, which means that S can be decomposed into a finite number of connected arcs, called skeleton branches, composed of points of degree two, and the branches meet at skeleton joints (or bifurcation points) that are points of degree three or higher." Skeletonization is used for optical character recognition [16], object matching and recognition [23], biomedical image analysis: vessel system geometrical and structural analysis and surgery planning [14, 17], lungs tree analysis [1], etc.

A Skeleton-based object descriptor aggregates geometry, symmetry and topology of its shape [9, 10]. Skeleton should contain the centers of maximal disks (medial axis points) lying inside of the object and touching boundaries at least in two points, which is used for object shape reconstruction. This shape descriptor is required to be invariant to translation, scale and rotation, since these transformations do not change the shape of the object. Demir *et al*. in [3] pointed out the main challenges of object skeletonization: dimensionality reduction while transforming the shape to the skeleton; the transition to continuous domain to get the best skeletal representation; the trade-of between skeleton simplicity and shape representational power.

There are three classical main ways of skeleton extraction: morphological thinning based on iterative boundary removal, geometric methods based on Voronoi diagram, distance transform based methods [4]. A good survey on skeletonization methods is provided by [16]. But such methods are sensitive to the noise on the boundary and require post-processing procedure for redundant branches pruning [16, 18].

According to the recent great success of convolutional neural networks in different computer vision tasks: classification, segmentation, object detection etc., their ability to represent data in the latent feature space could be used for direct skeleton extraction of the object without further pruning.

2. Related work

There are plenty of mathematical methods for skeleton extraction from the object shape [16], and a very few devoted to skeleton extraction based on Neural Networks. Skeletonization problem could be considered as a per-pixel classification problem known as semantic segmentation. This idea was adopted by Holistically-Nested Edge Detection (HED) method [22] – combination of fully convolutional network (FCN) and deep supervision. Authors referred to the preference of multiple scale predictions combination for final edge map generation. Most of the existing CNN-based architectures for real images skeletonization are based on HED architecture.

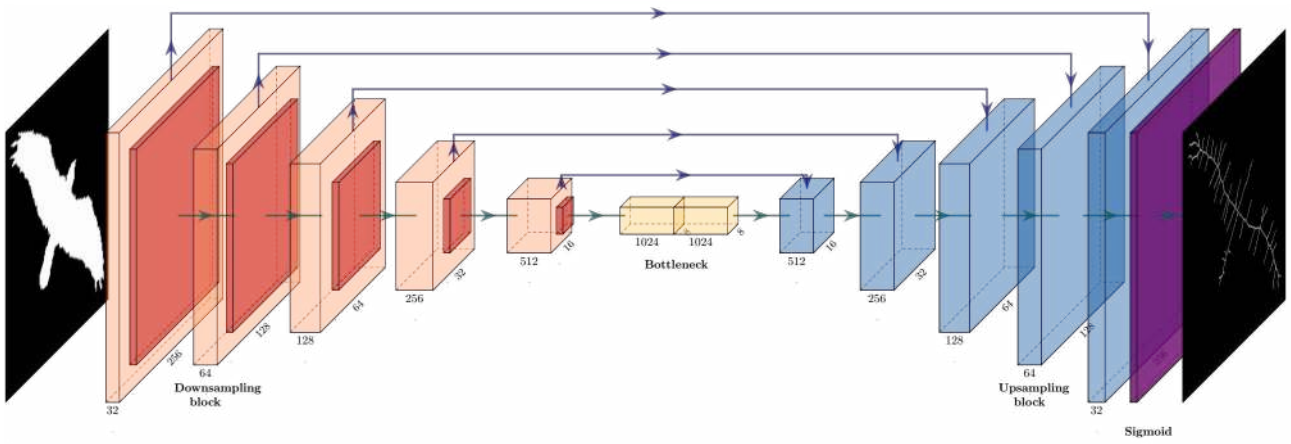

Figure 1. Modified U-Net architecture.

The first known CNN-based architecture for real image skeletonization is FSDS [19]. Side-output Residual Network (SRN) [8] overcomes a necessity of ground truth scale adjusting for each side-output via usage of Residual Units for side-outputs stacking. Linear Span Network follows the same idea as SRN with strong mathematical background explained the necessity of explicit connection between side-outputs, like RUs in SRN, on the base of Linear Span Theory [13]. Rich Side-output Residual Network [12] is a modification of SRN, which combines encoder block's hierarchical features in a special convolutional manner before passing them to the side-output. Similar idea is described in [25] where authors used hierarchical feature integration in their network (Hi-Fi) without any RUs. Multi-Scale Bidirectional FCN (MSB-FCN) [24], as Hi-Fi, described advantages of deep-to-shallow and shallow-to-deep features propagation and pyramid pooling usage instead of FCN-like side-outputs. Nowadays state-of-the-art model for skeletonization task on SK-LARGE Dataset named DeepFlux [20] is a compromise between classical skeletonization approaches and CNN-based one. It solved a skeleton extraction task as a pixel-wise regression task instead of classification.

All above methods inherited their architecture from FCN and were applied to skeleton extraction from real image task. There was no similar method for binary image skeletonization problem as well as a big enough dataset to solve it in Deep Learning fashion. In this paper we describe our CNN-based approach for object shape skeletonization within SkelNetOn 2019 competition [3]. We believe the modification of initial U-Net architecture [15], which is widely used basic architecture for semantic segmentation, could be used for skeletonization due to its ability to elegantly combine local well detailed features with context reach global features without deep supervision.

3. Methods

3.1. Modification of U-Net architecture

To solve the problem of skeleton extraction we started from vanilla U-Net architecture and modified it trying to maximize the model performance on a validation set. The model that allowed to get the highest scores on validation and test data depicted on Figure 1.

Each downsampling block is a sequence of five blocks that consist of convolution layer and a residual block. At the end of each downsampling block max pooling layer was applied. After five such blocks similarly to vanilla U-Net bottleneck was used. It contains two convolutional layers with 1x1 kernel and with ReLU activations after each. Upsampling blocks that follow after the bottleneck layers contain transposed convolution layer, concatenation with a corresponding output of downsampling block and four convolutional layers.

3.2. Weighed Focal Loss

To tackle class imbalance focal loss function [11] was modified to introduce a high penalty for incorrect misclassification of positive pixels:

$$FL = -w_{pos}(1-p)^{\gamma}log(p) - w_{neg}p^{\gamma}log(1-p),$$

where w_{pos} and w_{neg} - class weights for positive and negative class correspondingly, p - probability that sample belongs to positive class, γ - focusing parameter.

3.3. Data augmentations

In the case if dataset for training is quite small the data augmentation may be very important to improve performance and robustness of neural networks. Considering skeletonization problem, one should be very accurate with

data augmentation methods used, because some transformations of target image may lead to loss of some properties of resulting skeleton, that should be preserved. For example, scale and shear transformations may lead to a situation where skeleton is thicker or thinner than one pixel. Interpolation that is applied along with rotations that are not a multiple of 90 degrees may lead to a situation, where skeleton on target image becomes not invariant to rotations. That is why in our work we used only rotations on 90 degrees and horizontal flips of input images and targets.

4. Experimental results

The model was trained and evaluated on Pixel SkelNetOn 2019 Dataset [3]. The dataset contains 1725 black and white images with size 256x256 pixels. It was split into three subsets: training, validation and test subsets with 1219, 242 and 266 images correspondingly. To train the model we randomly split available training set on train and cross-validation in proportion 80:20. Cross-validation set was used for early stopping, reducing learning rate and saving model checkpoints. Adam optimizer was used and F1-score was used to evaluate the model performance.

In our experiments we have compared the binary cross entropy, weighted binary cross entropy, dice loss, Jaccard loss, focal loss, weighed focal loss and different combinations of listed above loss functions. The weighted focal loss showed the highest performance compared to other loss functions. A ratio between the number of background pixels and pixels that represent skeleton is near 127:1 in the training dataset. We compared different values of weights in the weighted focal loss and found that $w_{pos} = 50$ and $w_{neg} = 0.75$ along with $\gamma = 2$ give the best results.

Also, experiments with modifying residual blocks, adding squeeze excitation blocks [7] and attention layers [21] were conducted and we haven't seen any notable changes in model performance.

Experiments on validation data showed that the same model architectures trained on different subsets of training data may lead to notable differences in resulting predictions of the model - F1-score may vary from 0.6 to 0.84 and such score for each model was reproducible on validation and training sets. Two methods of combination results of inference were tested - a simple average of predictions and majority voting for each pixel. Averaging of predictions allowed to increase F1-score on 0.04 compared meanwhile voting algorithm did not produce any improvements.

We used cross-validation data to find a threshold that allows for producing the best binary outputs for each model. An average threshold was used to generate binary outputs for the whole ensemble and was equal to 0.81.

That is why we chose the best eight models for the final ensemble and averaged their predictions. Each model got F1-score 0.83-0.8415 on cross-validation subset from training data, 0.74-0.75 on validation data. The ensemble of eight models got F1-score 0.7846 on test data.

5. Discussion

A high deviation in model performance may be explained that models are sensitive to initialization and a set of samples used for training.

One of the features of the dataset is that it contains a lot of images with the same class in train, validation and test datasets. The difference between these images of the same class may be not big, but small variations may lead to notable differences in target images. That is why one of the open questions is how model performance would change for new classes that were not present within the training dataset.

During visual analysis of results of model inference, we noticed that in many cases it hard for the model to predict skeleton pixels on the intersection on skeleton lines. One of the ways for further improvement may be a usage of not pixel-wise loss functions that should evaluate the quality of skeletons and compliance with the requirements to skeletons in output images.

6. Conclusions

In this paper, we proposed a semantic segmentation network, that is based on a U-Net architecture and is able to tackle skeletonization problem when skeleton is represented as a binary image. Experiments were held to finetune network architecture and hyperparameters, find optimal loss function and an algorithm to merge model outputs into ensemble, which resulted in 0.75 F1-score on the validation set and the ensemble of eight identical models, trained on different data subsets, got 0.7846 F1-score on test data.

References

[1] Zijian Bian, Wenjun Tan, Jinzhu Yang, Jiren Liu, and Dazhe Zhao. Accurate airway centerline extraction based on topological thinning using graph-theoretic analysis. *Bio-medical materials and engineering*, 24(6):3239–3249, 2014.

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing.

[3] Ilke Demir, Camilla Hahn, Kathryn Leonard, Geraldine Morin, Dana Rahbani, Athina Panotopoulou, Amelie Fondevilla, Elena Balashova, Bastien Durix, and Adam Kortylewski. SkelNetOn 2019 Dataset and Challenge on Deep Learning for Geometric Shape Understanding. *arXiv e-prints*, 2019.

[4] D Ebert, P Brunet, and I Navazo. An augmented fast marching method for computing skeletons and centerlines. In *Proceedings of VisSym*, pages 251–258, 2002.

[5] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, Alexander Dovzhenko, Olaf Tietz, Cristina Dal Bosco, Sean Walsh, Deniz Saltukoglu, Tuan Leng Tay, Marco Prinz, Klaus Palme, Matias Simons, Ilka Diester, Thomas Brox, and Olaf Ronneberger. U-net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, 2019.

[6] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[8] Wei Ke, Jie Chen, Jianbin Jiao, Guoying Zhao, and Qixiang Ye. Srn: Side-output residual network for object symmetry detection in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[9] Joost Koehoorn, Cong Feng, Jacek Kustra, Andrei Jalba, and Alexandru Telea. Unified part-patch segmentation of mesh shapes using surface skeletons. In *Skeletonization*, pages 89–122. Elsevier, 2017.

[10] Louisa Lam, Seong-Whan Lee, and Ching Y Suen. Thinning methodologies-a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 14(9):869–885, 1992.

[11] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007, 2017.

[12] Chang Liu, Wei Ke, Jianbin Jiao, and Qixiang Ye. RSRN: rich side-output residual network for medial axis detection. In *ICCV Workshops*, pages 1739–1743. IEEE Computer Society, 2017.

[13] Chang Liu, Wei Ke, Fei Qin, and Qixiang Ye. Linear span network for object skeleton detection. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[14] KM Meiburger, SY Nam, E Chung, LJ Suggs, SY Emelianov, and F Molinari. Skeletonization algorithm-based blood vessel quantification using in vivo 3d photoacoustic imaging. *Physics in Medicine & Biology*, 61(22):7994, 2016.

[15] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (availa ble on arXiv:1505.04597 [cs.CV]).

[16] Punam K Saha, Gunilla Borgefors, and Gabriella Sanniti di Baja. A survey on skeletonization algorithms and their applications. *Pattern Recognition Letters*, 76:3–12, 2016.

[17] Dirk Selle, Bernhard Preim, Andrea Schenk, and H-O Peitgen. Analysis of vasculature for liver surgical planning. *IEEE transactions on medical imaging*, 21(11):1344–1357, 2002.

[18] Wei Shen, Xiang Bai, XingWei Yang, and Longin Jan Latecki. Skeleton pruning as trade-off between skeleton simplicity and reconstruction error. *Science China Information Sciences*, 56(4):1–14, 2013.

[19] Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Xiang Bai, and Alan Yuille. Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing*, 26(11):5298–5311, Nov 2017.

[20] Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, Sven Dickinson, and Kaleem Siddiqi. Deepflux for skeletons in the wild, 2018.

[21] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *Lecture Notes in Computer Science*, page 3–19, 2018.

[22] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *International Journal of Computer Vision*, 125(1-3):3–18, Mar 2017.

[23] Cong Yang, Oliver Tiebe, Kimiaki Shirahama, and Marcin Grzegorzek. Object matching with hierarchical skeletons. *Pattern Recogn.*, 55(C):183–197, July 2016.

[24] Fan Yang, Xin Li, Hong Cheng, Yuxiao Guo, Leiting Chen, and Jianping Li. Multi-scale bidirectional fcn for object skeleton extraction. In *AAAI*, 2018.

[25] Kai Zhao, Wei Shen, Shanghua Gao, Dandan Li, and Ming-Ming Cheng. Hi-fi: Hierarchical feature integration for skeleton detection. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul 2018.