Luca Rettenberger*, Marcel Schilling, and Markus Reischl

Annotation Efforts in Image Segmentation can be Reduced by Neural Network Bootstrapping

https://doi.org/10.1515/cdbme-2022-1084

Abstract: Modern medical technology offers potential for the automatic generation of datasets that can be fed into deep learning systems. However, even though raw data for supporting diagnostics can be obtained with manageable effort, generating annotations is burdensome and time-consuming. Since annotating images for semantic segmentation is particularly exhausting, methods to reduce the human effort are especially valuable. We propose a combined framework that utilizes unsupervised machine learning to automatically generate segmentation masks. Experiments on two biomedical datasets show that our approach generates noticeably better annotations than Otsu thresholding and k-means clustering without needing any additional manual effort. Using our framework, unannotated datasets can be amended with pre-annotations fully unsupervised thus reducing the human effort to a minimum.

Keywords: Semantic Segmentation, Machine Learning, Computer Vision, Image Annotation

1 Introdcution

In recent years, Deep Learning (DL) has advanced into nearly all areas of science with great success, medical imaging being no exception [1]. However, this theoretical potential often conflicts with the large effort needed to annotate data required for supervised learning. Especially in cases where segmentation masks have to be generated, domain experts need to spend a lot of valuable time annotating. This issue is particularly present in biomedical engineering where data is often ambiguous and complex [2, 3, 4]. For example, to automatically track laboratory animals [5] or biological cells [1], the position of each specimen needs to be annotated. Further, within the automatic quantification of objects in cancer diagnostics, a large number of samples need to be annotated to obtain sufficient results [3]. Consequently, data often remains unused, as domain experts cannot invest enough time to annotate them.

Several approaches can be taken to counteract this. Some solutions reduce the effort of pixel-wise annotations by requiring coarser descriptions than explicit segmentation masks [6, 7]. Such frameworks can reduce the effort required to annotate datasets considerably but still need human interaction for each sample. Additionally, the annotations are limited to the specific framework and cannot be used in general. Other methods try to reduce the needed human effort by using extended learning rules rather than explicit annotator support [8, 9, 10]. Those weakly-supervised models often do not perform satisfactorily, require large amounts of data, and complex additional training.

In contrast to previous methods, our framework requires minimal human interaction, generates complete segmentation masks, and does not require complicated additional steps. Further, empirical studies show that our method can be employed with smaller-sized datasets as they are common in biomedical engineering.

We propose a simple framework that generates high-quality segmentation masks which can be used as a pre-processing step to reduce human annotation effort. For this, we employ a simple process that autonomously activates a more complex one (bootstrapping). In our work, simple processes are methods that generate the segmentation masks by thresholding the intensity of pixel values and the complex process is a CNN. Through this interplay, spatial features are learned without human intervention and the CNN can generalize beyond the erroneous and noisy annotations provided by the simple processes. This is a sensible assumption, as it has been shown that moderate annotation noise does not harm the learning process [11]. The practitioner obtains a set of approximated annotations that can either be manually improved or directly employed in biomedical applications. We aim to relieve domain experts like pathologists from exhausting annotating labor and draw attention to the possibilities of automated dataset generation in the often small-scale biomedical domain.

Our method is open-source, ready to be employed in clinical practice, and available at https://github.com/Heterogeneous-Semantic-Segmentation/Reducing-Annotation-Efforts-by-DNN-Bootstrapping.

***Corresponding author: Luca Rettenberger,** Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, e-mail: luca.rettenberger@kit.edu

Marcel Schilling, Markus Reischl, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

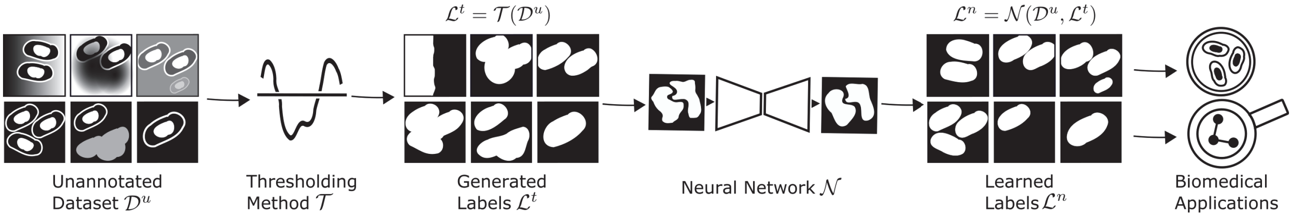
$\mathcal{L}^t = \mathcal{T}(\mathcal{D}^u)$
$\mathcal{L}^n = \mathcal{N}(\mathcal{D}^u, \mathcal{L}^t)$
Unannotated Dataset $\mathcal{D}^u$
Thresholding Method $\mathcal{T}$
Generated Labels $\mathcal{L}^t$
Neural Network $\mathcal{N}$
Learned Labels $\mathcal{L}^n$
Biomedical Applications

Fig. 1: Overview of our framework: An unannotated dataset $\mathcal{D}^u$ is provided to a thresholding method $\mathcal{T}$ which produces a corresponding set of segmentation masks $\mathcal{L}^t$. $\mathcal{L}^t$ is then employed into a deep neural network that can generalize beyond the annotations $\mathcal{L}^t$. This results in a improved set of segmentation masks $\mathcal{L}^n$ which can either be used to manually remove the remaining wrong annotations or directly be employed in biomedical applications.

2 Method

Figure 1 provides a conceptual overview of our developed framework. The main idea is to have some unannotated, raw, dataset $\mathcal{D}^u$ and a naïve process $\mathcal{T}$ which generates imperfect annotations $\mathcal{L}^t = \mathcal{T}(\mathcal{D}^u)$ due to its simplicity. $\mathcal{L}^t$ is then employed in a neural network $\mathcal{N}$ (or any more capable process than $\mathcal{T}$) to learn segmentation masks $\mathcal{L}^n = \mathcal{N}(\mathcal{D}^u, \mathcal{L}^t)$. Since we assume that $\mathcal{N}$ can learn correlations and is capable of generalization, it is expected that the noise from false annotations can be filtered out and the generated segmentation masks will be improved. $\mathcal{L}^n$ can then be used in a post-processing step to filter out insufficient labels or be further enhanced manually, depending on the requirements and quality of the generated dataset.

Even though we commit to a small subset of possible methods here, both $\mathcal{T}$ and $\mathcal{N}$ may be arbitrary processes that generate segmentation masks of different quality. In our case, the naïve method $\mathcal{T}$ is either the Otsu algorithm or k-means clustering of the intensity value of the respective image, where we expect the biggest cluster to be the background. In our work, $\mathcal{N}$ is always a U-Net [1], which is a deep, symmetric, fully convolutional neural network. It combines feature maps with (transposed) convolutional and pooling operations to generate segmentation masks. To evaluate the results, Dice-Sørensen Coefficient (DSC), Intersection over Union (IoU), and Pixel Accuracy (PA) are employed.

3 Results

3.1 Datasets

We evaluated our framework on two biomedical imaging datasets. First, we use the ISIC 2017 Melanoma image segmentation dataset introduced at the International Skin Imaging Collaboration (ISIC) challenge 2017 [3]. The dataset consists of 2000 train and 600 test samples each containing a binary segmentation mask describing the location of a skin lesion. Second, we consider the binary segmentation of spheroids in a high-throughput Droplet Microarray [4]. The 470 train and 118 test samples describe the location of the respective spheroid. Figure 2 shows two samples of the datasets.

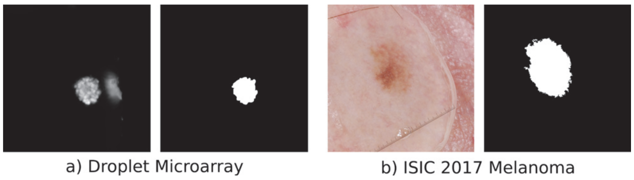

Fig. 2: Examples: One sample of each dataset used in this work including ground truth mask. a) High-throughput Droplet Microarray [4] and b) ISIC 2017 Melanoma image segmentation [3].

3.2 Architecture, Training, and Implementation

We used the Dice Loss as the objective function and the Adam optimizer with a learning rate of 0.001 in all experiments. All samples are resized to 256×256 pixels for both datasets. The following data augmentations are used for training: horizontal and vertical flipping, blurring, Gaussian noise, rotations, scaling, brightness variations, and contrast jittering. Details about the data augmentation are available in the project repository. We randomly divide the data with an 80% / 20% split into training and validation. The whole training dataset consists of the respectively generated segmentation masks provided by $\mathcal{T}$, only during test-time the actual annotations are available for evaluation. All samples are normalized to be in the range $[0, 1]$. The U-Net [1] is implemented in PyTorch Lightning. Apart from a simple extension to accept RGB images, the architecture is used in its original form. For the k-means algorithm, the implementation included in the OpenCV package is utilized and for Otsu thresholding scikit-image is used. The ISIC Melanoma images are converted to grayscale if used in the Otsu algorithm since it expects mono-channel images. The

Tab. 1: Evaluation results with respect to the evaluation metrics. The scores are calculated with the ground-truth segmentation masks as targets. U-Net(*X*) means that U-Net is trained with annotations provided by method *X*. Table a) shows the results for the Droplet Microarray dataset and Table b) for the ISIC Melanoma dataset. The standard deviation over four runs with different random seeds is given as ±. U-Net(Human) is the baseline performance, in which the ground truth annotations are provided for training.

	Otsu	**U-Net (Otsu)**	**k-means**	**U-Net (k-means)**	**U-Net (Human)**
DSC	0.34	0.76 ± 0.012	0.37	0.71 ± 0.012	0.93 ± 0.012
IoU	0.21	0.62 ± 0.001	0.23	0.56 ± 0.013	0.86 ± 0.022
PA	0.98	0.99 ± 0.001	0.98	0.99 ± 0.001	0.99 ± 0.001

a) Droplet Microarray

	Otsu	**U-Net (Otsu)**	**k-means**	**U-Net (k-means)**	**U-Net (Human)**
DSC	0.58	0.58 ± 0.027	0.54	0.64 ± 0.032	0.77 ± 0.012
IoU	0.42	0.45 ± 0.026	0.38	0.48 ± 0.037	0.63 ± 0.023
PA	0.84	0.84 ± 0.001	0.83	0.87 ± 0.002	0.93 ± 0.001

b) ISIC Melanoma

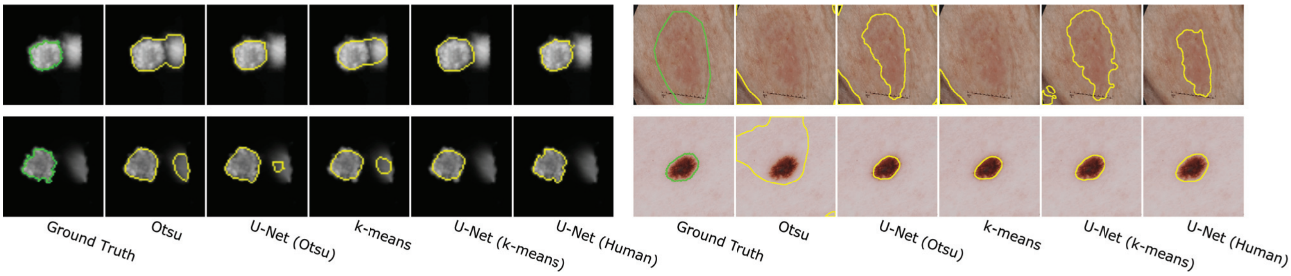

Fig. 3: Result samples: Two samples of both datasets with the learned segmentation masks. The annotations are given as contours over the respective sample. The ground truth is marked in green, the predicted masks in yellow.

training is performed on an NVIDIA GeForce RTX 3090 GPU and an AMD Ryzen 9 5950X 16-Core 3.40GHz CPU. We repeat all experiments four times with different random seeds to avoid initialization effects and provide mean metrics inducing standard deviation. A test split of the respective dataset is always used for the evaluation.

3.3 Experiments

Table 1 a) shows the results for the Droplet Microarray dataset. All metrics are computed with the actual annotations provided. The quality of the segmentation masks is improved by the U-Net in all cases. Otsu generates worse annotations (DSC = 0.34) compared to k-means (DSC=0.37). However, the U-Net seems to better generalize with the seemingly worse annotations of the Otsu algorithm (DSC=0.76) than it does with k-means (DSC=0.71). Nevertheless, with both datasets the U-Net produces noticeably better segmentation masks, as can be seen in Figure 3 on the left. In the upper sample, both thresholding methods fail to filter out the imaging reflection on the border. The U-Net trained with either of the generated labels recognizes and filters out those reflections. The lower sample also contains reflections which are not detectable by thresholding methods but are filtered out if the U-Net is trained and paired with k-means labels.

Table 1 b) displays the results for the ISIC Melanoma dataset. Improvements can also be detected here, but they are less pronounced than with the Droplet Microarray dataset. This is presumably since in this case, the background (human skin) is less homogeneous compared to the Droplet Microarray dataset and also contains RGB images. Both those circumstances make it more difficult for thresholding methods to segment the samples. With this dataset, the annotations provided by Otsu are slightly better (DSC=0.58) than k-means (DSC=0.54). Training U-Net with annotations generated by the Otsu algorithm does not yield much improvement (DSC=0.58). K-means seems to be more useful for training, as a larger boost in quality can be observed here (DSC=0.64). Figure 3 on the right displays two samples with the learned masks. For the upper sample, both Otsu and k-means generated incorrect annotations, however, the U-Net is much closer to the ground truth in both cases. The lower sample displays a wrong label for the Otsu algorithm which is visibly better after U-Net training. In this case, k-means already generates an appropriate label which is still slightly improved by the U-Net. The segmentation of lesions on the skin seems to be a generally difficult challenge since even with the ground-truth annotations given for training, the DSC is merely 0.77.

4 Discussion

With our framework, it is possible to reduce the manual labeling effort noticeably. This is especially valuable in the biomedical field, where domain experts have little time to spare and annotating is particularly complex. The core message is that a supervised method coupled with automatically generated annotations by a naïve process like Otsu or k-means can learn spatial features and determine correlations between the sample and its respective segmentation mask which reach beyond the provided annotations. Further, we show that extending the generation of pre-annotations with the help of deep neural networks can reduce the work of annotating without requiring additional effort since CNNs like U-Net are capable of learning associations that simple processes like thresholding cannot depict. However, in its current version, our framework will perform suboptimally on datasets where the class to be segmented is not clearly delineated. Yet, more complex datasets would also be solvable if the framework is extended. The results of our experiments suggest that annotating datasets without the help of a capable pre-annotation framework like ours is inefficient and should be avoided since adequate approximations of the desired annotations can be generated, which can considerably reduce the workload for domain experts when labeling datasets.

5 Conclusion

We recognize the problem that annotating datasets is very labor intensive and domain experts often lack the time, especially in the case of semantic segmentation. Therefore, we present a framework that automatically generates annotations. For this purpose, the interaction of simple thresholding methods and the U-Net architecture is used and investigated. The thresholding methods are simplistic and lead to erroneous and noisy annotations, which the U-Net employs to generalize beyond the noise and thus obtains much better segmentation masks. It is shown through two experiments with biomedical data that our framework is capable of generating useful annotations, which are not flawless but could speed up a potential annotation process considerably.

Future work could extend our framework to contain more steps with increasing annotation quality. Thus, an additional CNN might follow the U-Net. In addition, the threshold procedures may be extended by more complex processes to obtain the initial annotations in better quality and thus improve the proposal. Experiments focusing on the question of which annotations are helpful could also yield further insights, as our results show that the quantified quality of initial annotations does not necessarily correlate with the performance of the U-Net.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[2] Marcel P Schilling et al. “Label Assistant: A Workflow for Assisted Data Annotation in Image Segmentation Tasks.” In: *Proceedings-31. Workshop Computational Intelligence: Berlin, 25.-26. November 2021*. Vol. 25. KIT Scientific Publishing. 2021, p. 211.

[3] Noel CF Codella et al. “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi).” In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[4] Anna A Popova et al. “Facile one step formation and screening of tumor spheroids using droplet-microarray platform.” In: *Small* 15.25 (2019), p. 1901299.

[5] Roman Bruch et al. “epiTracker: A Framework for Highly Reliable Particle Tracking for the Quantitative Analysis of Fish Movements in Tanks.” In: *Slas Technology: Translating Life Sciences Innovation* 26.4 (2021), pp. 367–376.

[6] Di Lin et al. “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3159–3167.

[7] Jifeng Dai, Kaiming He, and Jian Sun. “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation.” In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1635–1643.

[8] Alexander Kolesnikov and Christoph H Lampert. “Seed, expand and constrain: Three principles for weakly-supervised image segmentation.” In: *European conference on computer vision*. Springer. 2016, pp. 695–711.

[9] Tianyi Zhang et al. “Decoupled spatial neural attention for weakly supervised semantic segmentation.” In: *IEEE Transactions on Multimedia* 21.11 (2019), pp. 2930–2941.

[10] Yunchao Wei et al. “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1568–1576.

[11] Marcel P Schilling et al. “Automated Annotator Variability Inspection for Biomedical Image Segmentation.” In: *IEEE access* (2022).