# Cost-efficient segmentation of electron microscopy images using active learning\*

Joris Roels<sup>1,2</sup>[0000-0002-2058-8134] and Yvan Saeys<sup>1,2</sup>[0000-0002-0415-1506]

Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

 ${\tt [jorisb.roels, yvan.saeys]@ugent.be}$ 

<sup>2</sup> Inflammation Research Center, Flanders Institute for Biotechnology, Ghent, Belgium

Abstract. Over the last decade, electron microscopy has improved up to a point that generating high quality gigavoxel sized datasets only requires a few hours. Automated image analysis, particularly image segmentation, however, has not evolved at the same pace. Even though state-of-the-art methods such as U-Net and DeepLab have improved segmentation performance substantially, the required amount of labels remains too expensive. Active learning is the subfield in machine learning that aims to mitigate this burden by selecting the samples that require labeling in a smart way. Many techniques have been proposed, particularly for image classification, to increase the steepness of learning curves. In this work, we extend these techniques to deep CNN based image segmentation. Our experiments on three different electron microscopy datasets show that active learning can improve segmentation quality by 10 to 15% in terms of Jaccard score compared to standard randomized sampling.

**Keywords:** Electron microscopy  $\cdot$  Image segmentation  $\cdot$  Active learning.

#### 1 Introduction

Semantic image segmentation, the task of assigning pixel-level object labels to an image, is a fundamental task in many applications and one of the most challenging problems in generic computer vision. Particularly in biomedical imaging such as electron microscopy (EM), where annotated data is very sparsely available and image data contains high resolution ( $\approx 5~\rm nm^3$ ) and ultrastructural content. Nevertheless, deep learning has caused significant improvements in this particular research domain, over the last years [6,11,8].

Even though the impressive advances that have been made so far, state-of-the-art techniques mostly rely on large annotated datasets. This is an impractical assumption and only satisfied for particular use-cases such as e.g. neuron segmentation [2]. For segmentation of alternative classes, research often falls back

<sup>\*</sup> We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. Y.S. is a Marylou Ingram Scholar.

to manual segmentation or interactive approaches that rely on shallow segmentation algorithms [14,3,1], which is costly or sacrifices performance.

This work focuses on active learning, a subdomain of machine learning that aims to minimize supervision without sacrificing predictive accuracy. This is achieved by iteratively querying a batch of samples to a label providing oracle, adding them to the train set and retraining the predictor. The challenge is to come up with a smart selection criterion to query samples and maximize the steepness of the training curve [13].

In this work, we employ state-of-the-art active learning approaches, commonly used for classification, to image segmentation. Particularly, we illustrate on three EM datasets that the amount of annotated samples can be reduced to a few hundreds to obtain close to fully supervised performance. We start by formally defining the active learning problem in the context of image segmentation in Section 2. In Section 3, we give an overview of commonly used, recent active learning approaches in classification [13] and how these techniques can be used in segmentation. This is followed by experimental results and a discussion in Section 4. Lastly, the paper is concluded in Section 5.

# 2 Notations

We consider the task of image segmentation, *i.e.* given an N pixel image  $\mathbf{x} \in X \subset \mathbb{R}^N$ , we aim to compute a pixel-level labeling  $\mathbf{y} \in Y$ , where  $Y = \{0, \dots, C-1\}^N$  is the label space and C is the number of classes. We particularly focus on the case of binary segmentation, *i.e.* C = 2. Let  $\mathbf{p}_j(\mathbf{x}) = [\mathbf{f}_{\theta}(\mathbf{x})]_j$  be the probability class distribution of pixel j of a parameterized segmentation algorithm  $\mathbf{f}_{\theta}$  (for example, an encoder-decoder network such as U-Net [11]).

Consider a large pool of i.i.d. sampled data points over the space  $Z = X \times Y$  as  $\{x_i, y_i\}_{i \in [n]}$ , where  $[n] = \{1, \ldots, n\}$ , and an initial pool of m randomly chosen distinct data points indexed by  $S_0 = \{i_j | i_j \in [n]\}_{j \in [m]}$ . An active learning algorithm initially only has access to  $\{x_i\}_{i \in [n]}$  and  $\{y_i\}_{i \in S_0}$  and iteratively extends the currently labeled pool  $S_t$  by querying k samples from the unlabeled set  $\{x_i\}_{i \in [n] \setminus S_t}$  to an oracle. After iteration t, the predictor is retrained with the available samples  $\{x_i\}_{i \in [n]}$  and labels  $\{y_i\}_{i \in S_t}$ , thereby improving the segmentation quality. Note that, without loss of generalization, the active learning approaches below are described for k = 1 as we can also query k > 1 samples for k iterations, without retraining. The complete active learning workflow is shown in Figure 1.

# 3 Active learning

In the following sections, we will discuss 5 well known and recent active learning approaches for classification: maximum entropy selection [9,10], least confidence selection [4], Bayesian active learning disagreement [7], k-means sampling [5] and core set active learning [12]. Furthermore, we will show how these techniques can be applied to image segmentation.

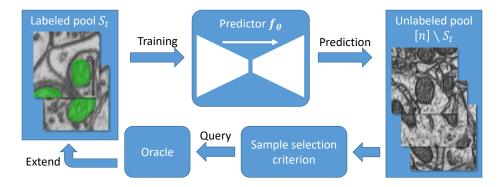


Fig. 1. Iterative active learning workflow for segmentation. A predictor network  $f_{\theta}$  predicts the class probability distributions of the unlabeled samples. These outputs are used in a sample criterion to select the 'most informative' samples. The selected samples are labeled by an oracle and the extended labeled pool is used to retrain the predictor.

#### 3.1 Maximum entropy sampling

Maximum entropy is a straightforward selection criterion that aims to select samples for which the predictions are uncertain [9,10]. Formally speaking, we adjust the selection criterion to a pixel-wise entropy calculation as follows:

$$x_{t+1}^* = \arg \max_{x \in [n] \setminus S_t} - \sum_{j=0}^{N-1} \sum_{c=0}^{C-1} [p_j(x)]_c \log [p_j(x)]_c.$$
 (1)

In other words, the entropy is calculated for each pixel and cumulated. Note that a high entropy will be obtained when  $p_j(x) = \frac{1}{C}$ , this is exactly when there is no real consensus on the predicted class (*i.e.* high uncertainty).

# 3.2 Least confidence sampling

Similar to maximum entropy sampling, the least confidence criterion selects samples for which the predictions are uncertain:

$$\boldsymbol{x}_{t+1}^* = \arg\min_{\boldsymbol{x} \in [n] \setminus S_t} \sum_{j=0}^{N-1} \max_{c=0,\dots,C-1} \left[ \boldsymbol{p}_j(\boldsymbol{x}) \right]_c. \tag{2}$$

As the name suggest, the least confidence criterion selects the probability that corresponds to the predicted class. Whenever this probability is small, the predictor is not confident about this decision. For image segmentation, we cumulate the maximum probabilities to select the least confident samples.

### 3.3 Bayesian active learning disagreement

The Bayesian active learning disagreement (BALD) approach [7] is specifically designed for convolutional neural networks (CNNs). It makes use of Bayesian CNNs in order to cope with the small amounts of training data that are usually available in active learning workflows. A Bayesian CNN assumes a prior probability distribution placed over the model parameters  $\theta \sim p(\theta)$ . The uncertainty in the weights induces prediction uncertainty by marginalising over the approximate posterior [7]:

$$[\mathbf{p}_j(\mathbf{x})]_c \approx \frac{1}{T} \sum_{t=0}^{T-1} \left[ \mathbf{p}_j(\mathbf{x}; \hat{\boldsymbol{\theta}}_t) \right]_c,$$
 (3)

where  $\hat{\boldsymbol{\theta}}_t \sim q(\boldsymbol{\theta})$  is the dropout distribution, which approximates the prior probability distribution p. In other words, a CNN is trained with dropout and inference is obtained by leaving dropout on. This causes uncertainty in the outcome that can be used in existing criteria such as maximum entropy (Equation (1)).

#### 3.4 K-means sampling

Uncertainty-based approaches typically sample close to the decision boundary of the classifier. This introduces an implicit bias that does not allow for data exploration. Most explorative approaches that aim to solve this problem transform the input x to a more compact and efficient representation z = g(x) (e.g. the feature representation before the fully connected stage in a classification CNN). The representation that we used in our segmentation approach was the bottleneck representation in the U-Net. The k-means sampling approach in particular then finds k clusters in this embedding using k-means clustering. The selected samples are then the k samples in the different clusters that are closest to the k centroids.

#### 3.5 Core set active learning

The core set approach [12] is a recently proposed active learning approach for CNNs that is not based on uncertainty or exploratory sampling. Similar to k-means, samples are selected from an embedding z = g(x) in such a way that a model trained on the selection of samples would be competitive for the remaining samples. Similar as before, the representation that we used in our segmentation approach was the bottleneck representation in the U-Net. In order to obtain such competitive samples, this approach aims to minimize the so-called core set loss. This is the difference between average empirical loss over the set of labeled samples (i.e.  $S_t$ ) and the average empirical loss over the entire dataset including unlabelled points (i.e. [n]).

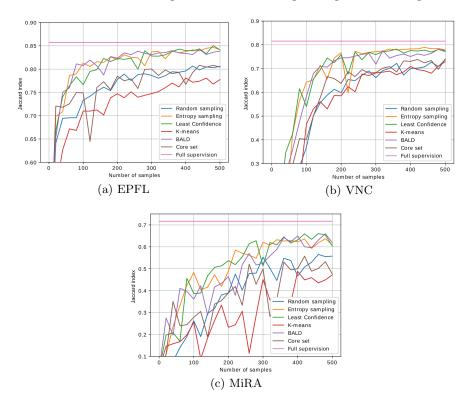


Fig. 2. Learning curves for the discussed active learning approaches for the different datasets.

#### 4 Experiments & discussion

Three public EM datasets where used to validate our approach:

- The EPFL dataset<sup>3</sup> represents a  $5 \times 5 \times 5 \mu m^3$  section taken from the CA1 hippocampus region of the brain, corresponding to a  $2048 \times 1536 \times 1065$ volume. Two  $1048 \times 786 \times 165$  subvolumes were manually labeled by experts for mitochondria. The data was acquired by a focused ion-beam scanning EM and the resolution of each voxel is approximately  $5 \times 5 \times 5$  nm<sup>3</sup>.
- The VNC dataset<sup>4</sup> represents two  $4.7 \times 4.7 \times 1 \mu m^3$  sections taken from the Drosophila melanogaster third instar larva ventral nerve cord, corresponding to a  $1024 \times 1024 \times 20$  volume. One stack was manually labeled by experts for mitochondria. The data was acquired by a transmission EM and the resolution of each voxel is approximately  $4.6 \times 4.6 \times 45 \text{ nm}^3$ .
- The MiRA dataset<sup>5</sup> [15] represents a  $17 \times 17 \times 1.6 \ \mu \text{m}^3$  section taken from the mouse cortex, corresponding to a  $8624 \times 8416 \times 31$  volume. The complete

<sup>&</sup>lt;sup>3</sup> Data available at https://cvlab.epfl.ch/data/data-em/

<sup>&</sup>lt;sup>4</sup> Data available at https://github.com/unidesigner/groundtruth-drosophila-vnc/ <sup>5</sup> Data available at http://95.163.198.142/MiRA/mitochondria31/

volume was manually labeled by experts for mitochondria. The data was acquired by an automated tape-collecting ultramicrotome scanning EM and the resolution of each voxel is approximately  $2 \times 2 \times 50 \text{ nm}^3$ .

To properly validate the discussed approaches, we split the available labeled data in a training and testing set. In the cases of a single labeled volume (VNC and MiRA), we split these datasets halfway along the y axis. A smaller U-Net (with 4 times less feature maps) was initially trained on m=20 randomly selected  $128 \times 128$  samples in the training volume (learning rate of  $1e^{-3}$  for 500 epochs). Next, we consider a pool of n=2000 samples in the training data to be queried. Each iteration, k=20 samples are selected from this pool based on one of the discussed selection criteria, and added to the labeled set  $S_t$ , after which the segmentation network is finetuned (learning rate of  $5e^{-4}$  for 200 epochs). This procedure is repeated for T=25 iterations, leading to a maximum training set size of 500 samples. We validate the segmentation performance with the well known Jaccard score:

$$J(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\sum_{i} [\boldsymbol{y} \cdot \hat{\boldsymbol{y}}]_{i}}{\sum_{i} [\boldsymbol{y}]_{i} + \sum_{i} [\hat{\boldsymbol{y}}]_{i} - \sum_{i} [\boldsymbol{y} \cdot \hat{\boldsymbol{y}}]_{i}}$$
(4)

This segmentation metric is also known as the intersection-over-union (IoU).

The resulting performance curves of the discussed approaches on the three datasets are shown in Figure 2. We additionally show the performance obtained by full supervision (i.e. all labels are available during training), which is the maximum achievable segmentation performance. In comparison to the random sampling baseline, we observe that the maximum entropy, least confidence and BALD approach perform significantly better. These methods obtain about 10 to 15% performance increase for the same amount of available labels for all datasets. The recently proposed core set approach performs similar to slightly better than the baseline. We expect that this method can be improved by considering alternative embeddings. Lastly, we see that k-means performs significantly worse than random sampling. Even though this could also be an embedding problem such as with the core set approach, we think that exploratory sampling alone will not allow the predictor to learn from challenging samples, which are usually outliers. We expect that a hybrid approach based on both exploration and uncertainty might lead to better results, and consider this future work.

Figure 3 shows qualitative segmentation results on the EPFL dataset. In particular, we show results of the random, k-means and maximum entropy sampling methods using 120 samples, and compare this to the fully supervised approach. The maximum entropy sampling technique is able to improve the others by a large margin and closes the gap towards fully supervised learning significantly.

Lastly, we are interested in what type of samples the active learning approaches select for training. Figure 4 shows 4 samples of the VNC dataset that correspond to the highest prioritized samples, according to the least confidence criterion, that were selected in the first 4 iterations. The top row illustrates the probability predictions of the network at that point in time, whereas the bottom row shows the pixel-wise uncertainty of the sample (*i.e.* the maximum in

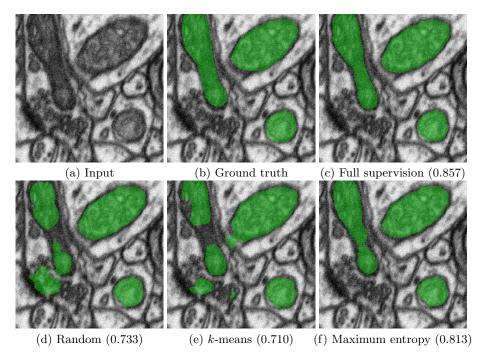
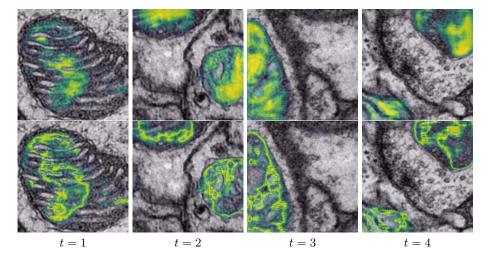


Fig. 3. Segmentation results obtained from an actively learned U-Net with 120 samples of the EPFL dataset based on random, k-means and maximum entropy sampling, and a comparison to the fully supervised approach. Jaccard scores are indicated between brackets.

Equation (2)). Note that the initial predictions at t=1 are of poor quality, as the network was only trained on 20 samples. Moreover, the uncertainty is high in regions where the network is uncertain, but it is low in regions where the network is wrong. The latter is a common issue in active learning and related to the exploration vs. uncertainty trade off. However, over time, we see that the network performance improves and more challenging samples are being queried to the oracle.

# 5 Conclusion

Image segmentation is one of the most challenging computer vision tasks, particularly for biomedical data such as electron microscopy as annotations are sparsely available. In order to be practically usable and scalable, image segmentation algorithms such as U-Net need to be able to cope with smaller amounts of annotated data. In this work, we propose to employ recent active learning approaches to minimize annotation efforts for training segmentation networks. Specifically, several of these approaches (e.g. maximum entropy and least confidence sampling) obtain the same performance as the random sampling baseline,



**Fig. 4.** Illustration of the selected samples in the VNC dataset over time in the active learning process. The top row shows the pixel-wise prediction of the selected samples at iterations 1 through 4. The bottom row show the pixel-wise least confidence score on the corresponding images.

but require 4 times fewer annotations. In future work, we will further minimize labeling efforts, by combining this active learning paradigm with weakly supervised approaches (*i.e.* using partially annotated data).

# References

- Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K.W., Schindelin, J., Cardona, A., Seung, H.S.: Trainable Weka Segmentation: A machine learning tool for microscopy pixel classification. Bioinformatics (2017). https://doi.org/10.1093/bioinformatics/btx180
- Arganda-Carreras, I., Turaga, S.C., Berger, D.R., Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., Laptev, D., Dwivedi, S., Buhmann, J.M., Liu, T., Seyedhosseini, M., Tasdizen, T., Kamentsky, L., Burget, R., Uher, V., Tan, X., Sun, C., Pham, T.D., Bas, E., Uzunbas, M.G., Cardona, A., Schindelin, J., Seung, H.S.: Crowdsourcing the creation of image segmentation algorithms for connectomics. Frontiers in Neuroanatomy 9 (2015). https://doi.org/10.3389/fnana.2015.00142, http://journal.frontiersin.org/Article/10.3389/fnana.2015.00142/abstract
- 3. Belevich, I., Joensuu, M., Kumar, D., Vihinen, H., Jokitalo, E.: Microscopy Image Browser: A Platform for Segmentation and Analysis of Multidimensional Datasets. PLoS Biology 14(1) (2016). https://doi.org/10.1371/journal.pbio.1002340
- 4. Blinker, K.: Incorporating Diversity in Active Learning with Support Vector Machines. In: Proceedings, Twentieth International Conference on Machine Learning (2003)
- 5. Bodo, Z., Minier, Z., Csato, L.: Active Learning with Clustering. Active Learning and Experimental Design @ AISTATS (2011)

- Ciresan, D.C., Giusti, A., Gambardella, M., L., Schmidhuber, J.: Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. NIPS pp. 1–9 (2012)
- Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: International Conference on Machine Learning (2017)
- 8. Januszewski, M., Kornfeld, J., Li, P.H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., Jain, V.: High-precision automated reconstruction of neurons with flood-filling networks. Nature Methods (2018). https://doi.org/10.1038/s41592-018-0049-4
- Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
- 10. Li, X., Guo, Y.: Adaptive Active Learning for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 859–866 (2013)
- Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention MICCAI 2015 pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4-28
- Sener, O., Savarese, S.: Active Learning for Convolutional Neural Networks: A Core-Set Approach. In: International Conference on Learning Representations (2018)
- 13. Settles, B.: Active Learning Literature Survey. Tech. rep., University of Wisconsin (2010). https://doi.org/10.1.1.167.4245
- 14. Sommer, C., Straehle, C., Kothe, U., Hamprecht, F.A.: Ilastik: Interactive learning and segmentation toolkit. In: IEEE International Symposium on Biomedical Imaging. pp. 230–233 (2011). https://doi.org/10.1109/ISBI.2011.5872394
- 15. Xiao, C., Chen, X., Li, W., Li, L., Wang, L., Xie, Q., Han, H.: Automatic Mitochondria Segmentation for EM Data Using a 3D Supervised Convolutional Network. Frontiers in Neuroanatomy (2018). https://doi.org/10.3389/fnana.2018.00092