Cell Detection With Deep Convolutional Networks Trained With Minimal Annotations


Published in:
IEEE International Symposium on Biomedical Imaging (ISBI) 2019 proceeding

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2018 IEEE
This work is made available online in accordance with the publisher’s policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date: 04. Feb. 2021
ABSTRACT

In this paper, we propose and thoroughly compare strategies for training convolutional neural networks (CNNs) for cell localization and segmentation in microscopy images with both little training data and in presence of significant label noise. Insufficient availability of ground truth (GT) is a common issue in the field of microscopy image analysis, hence the usefulness of such approaches. Performance evaluation is done using phase contrast microscopy human fibrosarcoma (HT1080) cells and comparing the resulting F-scores.

Index Terms—Cell detection, U-Net, convolutional neural networks, phase contrast microscopy

1. INTRODUCTION

A significant amount of biological research and clinical trials rely on the ability to accurately detect and/or segment cells in microscopy images. Meticulous tasks such as precise cell segmentation, localization, and counting are slow and error-prone if done manually by the medical researcher, which is why microscopy image analysis has increasingly relied on automated techniques [1]. Challenges associated with automated cell detection include their constant shape variation, frequent overlapping (especially in high density situations), and the presence of background noise. These difficulties haven’t been completely solved [2], although many promising results have recently been achieved, in particular thanks to advances in the field of deep learning (DL) [3, 4].

Some of the earlier CNN methods, such as [5], used a sliding window technique to segment cells on a pixel by pixel basis. More recently, the best cell segmentation results have been achieved by approaches based on either the U-Net [6] or Mask R-CNN [7]. The former makes use of an encoder-decoder architecture to predict the full image detection map, while the latter first generates region proposals, and then predicts the masks and classes. Their better use of the spatial information in the full image is an improvement over methods based on image crops, especially when cells cluster or overlap.

While DL approaches provide state of the art results in cell detection, they rely on the availability of large annotated datasets. This assumption is however uncommon for microscopy images, where datasets are rarely annotated completely or in detail, but instead, only a few interesting cells on a few images are ground truthed by the pathologist. This lack of reliable training severely hampers or even makes impossible the application of deep learning and CNN approaches [8]. Few papers have dealt with these circumstances. [9] recently proposed a U-Net based approach for limited training medical imaging data. However, it was only designed and tested for simple circular cell shapes, mostly performing localization rather than precise segmentation and assuming perfect GT is available for at least hundreds of training images. In this work, we compare several kinds of training approaches: using a semi- or fully annotated training set, with or without data augmentation, for localization or segmentation, with a more or less limited amount of training images. We thus analyse the influence of incomplete data, a more restricted amount of images (as little as 12), and complex and realistic cell shapes.

Fig. 1: U-Net architecture. Thickness represents the number of channels, height the size of the feature maps.
2. METHODS

We consider several training strategies for dealing with incomplete, inaccurate, and extremely limited data, all based on a U-Net architecture [6]. U-Net has been chosen since it has been proved effective even with few training images [9, 10], which makes it a very valuable tool for biomedical image segmentation.

2.1. U-Net architecture

Fig. 1 depicts our specific U-Net architecture used in this paper. It consists of two main parts: an encoder and a decoder. The encoder part ensures that context is captured and downsamples the feature maps at each layer, while increasing their number to extract discriminative features that are useful for cell detection. The decoder part upsamples the feature maps symmetrically, so the localization/segmentation of the cell is reconstructed at the original image scale. Skip connections [11], shown as purple arrows, allow recovering spatial information, lost during downsampling from the high-resolution decoder feature, which can be crucial for the localization task.

A block of the encoder consists of two consecutive $3 \times 3$ convolutions, each followed by a batch normalization, followed by a rectified linear unit (ReLU). After every block, $2 \times 2$ maxpooling with a stride 2 is applied, thus downsampling the features by half.

A block of the decoder similarly consists of two convolutions, batch normalizations, and ReLU, but is preceded by a $2 \times 2$ transposed convolution, in order to upsample the features to double the size. The last layer of the U-Net consists of a $1 \times 1$ convolution, followed by a sigmoid function in order to regress the probability of each pixel to belong to a cell (segmentation task) and/or estimate the distance to the cell centre (localization task).

2.2. Training

All networks described were implemented on PyTorch and trained using stochastic gradient descent (SGD), with a weight decay of 0.0005, a momentum of 0.9, and a learning rate of 0.1. The batch size is set to 1. We made use of two different loss functions. When training a U-Net for segmentation, we used binary cross-entropy (BCE) loss, which is standard for binary classification and given by $L_{\text{BCE}} = - (y \log(p) + (1-y) \log(1-p))$, where $p$ is the predicted probability and $y$ is the pixel label, i.e. $y = 1$ for pixels within a cell and $y = 0$ otherwise. We chose BCE over Dice loss because it is smoother, thus allowing for more stable training. When training it for localization, we used mean squared error loss $L_2$.

We propose several training strategies, detailed in 3.3, for generating cell detection models under the constraints of limited and/or incomplete annotations. Thus, we propose the use of efficient GT annotation using only cell centroid locations, represented by white Gaussian spots on a black background, versus the full cell masks. While less suited for accurate segmentation, such simple representation is the most commonly available and can be automatically generated from the kind of annotation medical experts usually provide. Moreover, we compare the strategy of using more training images but noisy or incomplete versus limited but perfectly annotated data, both from scratch or as a fine tuning strategy on pre-trained imperfect models. Finally, we also evaluate the influence of data augmentation (DA) as usual technique to mitigate the lack of data under this incomplete annotation scenario.

3. EXPERIMENTAL RESULTS

3.1. Dataset

The data used for training and testing was acquired in the context of 2D chemotaxis assays and consists of high resolution $2048 \times 1532$ phase contrast microscopy images of HT1080 cells. Images were captured using a Nikon Eclipse Ti-E microscope with a $4 \times$ objective. The scale is 1.61 μm/pixel, and cell density varies over the different images.

This dataset is composed of a total of 2465 such images, among which 2241 have partial and largely incomplete annotations provided by biologists for their own research. On average, only two thirds of the cells present in each image were annotated, with this percentage decreasing drastically for high density images. Moreover, the annotations only consist of the cell center locations as it is usual practice in most biological experiments. Another 212 were more thoroughly
labelled manually for centroid locations by ourselves, and are exclusively used as testing set to validate our approaches. Finally, we manually drew masks for the remaining 12 images, to be used for exploring the precise cell segmentation capabilities of our approaches. Fig. 2 shows an example of these different annotation types.

3.2. Preprocessing and data augmentation

Each image is preprocessed using the following steps. The image is converted to grey level, and a $1000 \times 2000$ center crop is made in order to eliminate the areas that are outside of the region of interest. This crop is then downsampled by a factor 4 and split into two squares of $250 \times 250$. Finally, both squares are converted to tensors, normalized, and each one is passed into the network along with the corresponding labelled image in case of training.

For some of the experiments, we also apply DA. As augmentation techniques, three squares of $250 \times 250$ are randomly cropped for each image, as well as their vertically and horizontally flipped versions.

3.3. Setup

We consider different approaches, which we are describing below, that combine the different training strategies to deal with the incomplete and/or small training sets explained in section 2.1.

Version A: The network is trained over 10 epochs with all the 2241 available partially labelled images. The GT is represented by Gaussian spots at the locations given by the partial manual annotations (see Fig. 2 (b)). This version is tested with and without DA as described in 3.2.

Version B: The network is first trained as in Version A (with and without DA). In order to fine-tune it to improve their noisy initial data, it is then further trained over 50 epochs on the reduced accurate training set consisting of the 12 cleanly labelled images.

Version C: The network is trained for segmentation over 50 epochs, using exclusively the 12 perfect masked images. GT image is represented as the full segmented cell masks (see Fig. 2 (c)).

Version D: The same training strategy as in Version C is used, but the GT image is converted into localization information as versions A and B, by drawing Gaussian spots at the segmented objects’ centroid locations (see Fig. 2 (d)).

The proposed approaches are compared against the following two baseline approaches:

Imaris: The cell segmentation feature of this extensively used software (Imaris v 9.2.1, Bitplane AG) was applied, involving smoothing followed by automatic thresholding. Batch parameters were empirically optimised from segmenting one image as well as possible, and saved for processing the entire dataset.

Convolutional Network approach using sliding window (CNN-SW): A small CNN was trained, taking $100 \times 100$ pixel crops around each annotated cell in our partially labelled dataset as input and classifying it as containing a cell or not. Random crops over the image that do not overlap any GT are used as examples of non-cell locations. This cropping training strategy allows for more than 20000 training crops in our 2241 partially labelled dataset, which are enough for a medium CNN architecture. It also mitigates the influence of incomplete annotation, since those are only occasionally introduced as non-cell examples. For testing, a sliding window approach is used to generate the final binary detection over the full image. This CNN has two convolution layers with max pooling and ReLU, followed by three fully connected layers. The network was trained over 30 epochs, using cross-entropy loss and SGD with a learning rate of 0.001 and a momentum of 0.9. The batch size was 4.

3.4. Postprocessing

After generating a prediction, a few more steps are required in order to extract the predicted cell locations from the output image from the network (see Fig. 3) and compare it with the GT. A binary threshold $t = 0.5$ is applied to the prediction, followed by the computation of contour moments using OpenCV’s contour features. From that we extract the corresponding centroids. These steps are applied for all versions except C, which already predicts a binary image and whose output can be used directly as segmentation result or used for localization by simply applying the centroid coordinates extraction step, and for CNN-SW, where we make use of non maxima suppression.

3.5. Results

A quantitative performance comparison was carried out between our proposed approaches and the previously explained state-of-art ones. Detection performance was quantitatively evaluated by computing true positives TP, false positives FP, false negatives FN, mean Euclidean distance $d$ from a GT cell to its associated TP detection centre, and the F-score $f = 2TP/(2TP + FP + FN)$. Matching between detection and GT is calculated by solving the linear sum assignment problem [12] with a cut-off value of 70 for the Euclidean distance between their object centers. This ensures that, in case of several detections matched to the same GT, only one is considered as TP while the rest are labeled as FP.

Among the proposed methods, the best performance in terms of F-score is achieved by B since it achieved the best compromise between TP, FP and FN. This validates the strategy of using as much data as available, even if it is inaccurate or incomplete, but only if some perfect annotation can be used for fine tuning as a reinforcement learning strategy. Other interesting conclusions can be extracted from the results in Table 1. The fact that both B versions respectively
outperform the A versions shows the usefulness of fine tuning with a smaller but cleaner training set, but also that this is likely to result in more cell detections but at the cost of more false detections. That needs to be considered depending on the final goal of specific biomedical applications. The comparison of the augmented versions with the non augmented versions shows that data augmentation doesn’t appear to provide a performance improvement. A likely explanation for this behaviour may be the amplification of label noise when augmenting the inaccurate images.

Despite the use of a very small training set, methods C and D achieve a good F-score, only slightly lower than the best method, and the most precise localization. D also happens to achieve the lowest number of FNs and best detection rate, since it has not been exposed to inaccurate data. This validates and the potential of U-net for dealing with extremely limited available data if this data is accurate.

We can then conclude that, for those applications where detection rate is more important than false detections, the use of perfectly annotated images (either in isolation or for fine tuning) is crucial. However, for the applications where false alarms are not allowed, it may be better to use large datasets despite their inaccuracies or incompleteness.

All our approaches greatly outperform the non learning based cell detection feature of Imaris, which tends to oversegment the cells and create huge amount of false detections, as well as the classical CNN-SW, where their cropping strategy tends to underdetect, limiting the detections to mostly clear and isolated cells.

### Table 1: Detection errors and F-score for the different approaches for cell localization.

<table>
<thead>
<tr>
<th>Method</th>
<th>GT</th>
<th>Metrics</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>d</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>×</td>
<td>Partial</td>
<td>13483</td>
<td>3444</td>
<td>6835</td>
<td>26.6</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>Perfect</td>
<td>14974</td>
<td>6246</td>
<td>5344</td>
<td>24.2</td>
<td>0.72</td>
</tr>
<tr>
<td>B</td>
<td>×</td>
<td>Augment</td>
<td>17557</td>
<td>7381</td>
<td>2761</td>
<td>22.1</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>Fine-1</td>
<td>16412</td>
<td>8028</td>
<td>3906</td>
<td>22.7</td>
<td>0.73</td>
</tr>
<tr>
<td>C</td>
<td>×</td>
<td></td>
<td>16969</td>
<td>7420</td>
<td>3392</td>
<td>17.8</td>
<td>0.76</td>
</tr>
<tr>
<td>D</td>
<td>×</td>
<td></td>
<td>17704</td>
<td>8519</td>
<td>2614</td>
<td>19.6</td>
<td>0.76</td>
</tr>
<tr>
<td>CNN-SW</td>
<td>×</td>
<td></td>
<td>9665</td>
<td>5773</td>
<td>10653</td>
<td>36.28</td>
<td>0.55</td>
</tr>
<tr>
<td>Imaris</td>
<td>×</td>
<td></td>
<td>19395</td>
<td>37051</td>
<td>923</td>
<td>16.19</td>
<td>0.51</td>
</tr>
</tbody>
</table>

In this paper, we proposed and evaluated different strategies for training a convolutional neural network for cell detection in phase contrast microscopy under conditions where very little and or inaccurate training data is available. Using those strategies, we were able to provide an efficient cell detection system that can be trained with as little as 12 images, and to improve models in presence of incomplete or inaccurate annotations. Our experiments showed that the approaches proposed easily outperform traditional non learning based approaches in terms of F-score, as well as learning based CNN approaches that are designed for big and accurate supervised datasets which are not common in the medical field.

### 4. CONCLUSION

This work has received funding from the EU’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 642866. We thank Lea Tomášová and ibidi GmbH for providing the data.

### 5. ACKNOWLEDGEMENTS
6. REFERENCES


