Insights into the behaviour of multi-task deep neural networks for medical image segmentation.

Bienias, Lukasz Tomasz; Rubio Guillamón, Juan José; Nielsen, Line Hagner; Alstrøm, Tommy Sonne

Published in:

Link to article, DOI:
10.1109/MLSP.2019.8918753

Publication date:
2019

Document Version
Peer reviewed version

Citation (APA):
ABSTRACT

Glandular morphology is used by pathologists to assess the malignancy of different adenocarcinomas. This process involves conducting gland segmentation task. The common approach in specialised domains, such as medical imaging, is to design complex architectures in a multi-task learning setup. Generally, these approaches rely on substantial post-processing efforts. Moreover, a predominant notion is that general purpose models are not suitable for gland instance segmentation. We analyse the behaviour of two architectures: SA-FCN and Mask R-CNN. We compare the impact of post-processing on the final predictive results and the performance of generic and specific models for the gland segmentation problem. Our results highlight the dependency of post-processing on tailored models as well as comparable results when using a generic model. Thus, in the interest of time, it is worth considering to use and improve generic models as opposed to design complex architectures when tackling new domains.

Index Terms—model optimisation, multi-task learning, image segmentation, deep learning, convolutional neural networks

1. INTRODUCTION

Recent developments in machine learning, especially deep learning, have led to a rapid growth in applications in the field of medical image processing [1]. For example, glandular morphology is used by pathologists to assess the malignancy of different adenocarcinomas, including prostate, breasts or colon, and this process involves analysing histological images in order to accurately segment the glands (examples shown in Fig. 1). Manual handling of these activities, usually carried out by pathologists, has many disadvantages, such as limited reproducibility or high time consumption. For these reasons, the automation of segmentation methods would be very beneficial for both physicians and patients. A challenge in gland segmentation is to have them detected and delimited individually, which we call a task of instance segmentation.

Xu et al. stated that glands segmentation is more demanding than recognising natural images (e.g. COCO dataset) [3]. This is mainly due to three facts. Firstly, the mathematical shape models have great difficulties in handling heterogeneous shapes of glands [4]. Secondly, variability of intracellular and extra-cellular matrices very often leads to anisochromasia [3] (the non-uniform colour of erythrocytes caused by a non-uniform distribution of haemoglobin in them). As a consequence, the background part of histopathological image contains more noise, such as the intensity gradients, than average natural images. Thirdly, many objects are closely located, which is known as coalescence problem and poses a great challenge in separating instances.

Deep learning has revolutionised computer vision by introducing the end-to-end learning, or representation learning, where the neural network learns to discover the underlying patterns (features) without explicit feature engineering. Currently, the most common approach for the instance segmentation of medical images is a domain specific neural net-
We use multi-task learning (MTL), where the main goal is to improve learning efficiency as well as prediction accuracy [5,6].

The 2015 MICCAI Gland Cell Challenge, won by Chen et al. [7], drew attention to the gland segmentation problem [2]. The objective of the challenge was to develop a system for instance segmentation of Haematoxylin and Eosin (H&E) stained slides images, Fig. 1 [8]. In 2016, Xu et al. proposed a new method that surpassed the winners of the challenge, where a multi-task network with three tasks was used: foreground segmentation, edge and object detection. In 2017, Yang et al. [9] described a Suggestive Annotation Fully Convolutional Network (SA-FCN), which attained SOTA for this dataset. More details on the SA-FCN will be given in Section 2. The SA-FCN is treated as representative of a domain specific system in this paper. All the aforementioned papers emphasise the beneficial effect of MTL but none of them addressed the impact of MTL on the individual task performance, which will help to understand how the individual tasks are sharing information between each other.

Besides, the common denominator of all these systems is that the architectures themselves solve only the task of semantic segmentation. Post-processing methods not only allow to solve the problem of the instance segmentation, but also enables to hide imperfections of the architecture output in a great extent. In many publications, the importance of post-processing activities seems to be neglected, since the process is usually poorly described. Moreover, by posting only results after post-processing, it is difficult to assess how well the architecture has been trained, since post-processing can hide a lot of defects of the network. This leads to the conclusion that it would be worth publishing the results before and after post-processing, which would give the reader a better picture of the system.

Xu et al. believe that general purpose models are not suitable for gland instance segmentation, only tailored systems [3]. We are challenging this claim, by demonstrating that “general purpose” image segmentation algorithm, such as the Mask R-CNN [10], can obtain similar performance. The Mask R-CNN model is considered as one of the best architectures for instance segmentation of nature images [11]. Moreover, we are showing that generic networks are not so sensitive to post-processing operations as the domain specific neural networks.

Our major contributions are: 1) we demonstrate deep models can be highly sensitive to post-processing methods, 2) we demonstrate that generic architectures such as Mask R-CNN obtains comparable results to domain designed models, which is in contrary to previous claims [3], where it is argued that “general purpose” models are not suitable for gland instance segmentation.

2. METHODS

2.1. SA-FCN Architecture

The SA-FCN consists of one encoder and two decoders, one for the foreground recognition and one for edge detection, see Fig. 2. The encoder is strongly inspired by the DCAN structure [7]. It consists of input block with two convolutional layers and five ”Intermediate Feature Maps” structures. Each ”Intermediate Feature Map” consists of two bottleneck blocks (inspired by the residual network [12]) and max pooling layer. The bottleneck design allows to reduce number of parameters, while maintaining a similar number of feature channels at the end of each residual module [9].

The decoding part of the network is a modified structure of Deeplab design [13], where the size of feature maps is gradually increased, to ensure smooth results [9]. The softmax function is used in the final layer of the network, to provide the classification scores. Model loss is defined as softmax cross entropy loss for both, the contour detection and foreground detection. The final loss is a weighted sum of foreground and contour losses and is formulated as $L_{\text{final}} = \lambda L_{\text{foreground}} + (1 - \lambda)L_{\text{contour}}$.

To increase the robustness and avoid overfitting, authors of the SA-FCN are using data augmentation transformations: random flipping on both axis, rotation by random multiples of 90 degrees and random $384 \times 384$ cropping.

To generate the final instance segmentation results, post-processing method is applied. Predicted probability maps of contour are subtracted from gland objects and post-processing steps are performed, including: dilation with a disk filter (radius 2) and removing spurious segments. Finally, each connected component is labelled with a unique value to represent one segmented gland. More details can be found in [9].

The DCAN post-processing methodology is similar to the SA-FCN. Nevertheless, the fusion of probability maps of contours and objects is slightly different and post-processing steps contain smoothing with a disk filter (radius 3), filling holes and removing small areas. More details about subsequent steps can be found in [7]. This method will be also tested on our implementation output.
2.2. Mask R-CNN Architecture

The Mask R-CNN is considered as one of the best model for object instance segmentation of natural images, announced as a the state-of-the-art for COCO dataset by [10]. It is conceptually simple and flexible framework, which detects objects in an image and simultaneously generates high-quality segmentation mask for each instance.

Mask R-CNN is a two stage model built upon Faster R-CNN [14] (presented in the Fig. 3). The first stage scans the image and generates proposals (areas likely to contain an object). The second stage classifies the proposals and generates bounding boxes and masks. More detailed description of the model can be found in [10]. To the best of out knowledge, no tests results for the Mask R-CNN on the 2015 MICCAI Gland Challenge dataset are published.

2.3. Our post-processing method

From the experimental results, as shown in Section 3.3, it was observed that post-processing strategies, proposed by SA-FCN and DCAN authors, have some limitations to be addressed, the elimination of which can positively affect the final result of the instance segmentation. For this reason, a modified post-processing method was developed, which consists of the following steps:

1. The output of the SA-FCN network, which are predicted probability maps of contours \( p_c \{ x; W_s, W_c \} \) and objects \( p_f \{ x; W_s, W_f \} \) respectively, are fused to generate a final segmentation masks by:

\[
m(x) = \begin{cases} 
1 & \text{if } p_f \{ x; W_s, W_f \} - p_c \{ x; W_s, W_c \} \geq 0.5, \\
0 & \text{otherwise.}
\end{cases}
\]

2. Small spurious segments are removed. Then the opening operation (erosion followed by dilation) is conducted. Subsequently, the holes in the prediction are filled.

3. Connected-component labelling method is utilised to represent individual segmented objects.

The main differences in relation to the SA-FCN method are the introduction of the opening operation, filling holes and slightly changing the order of operations. In our case, small spurious segments are removed immediately after the probability maps fusion.

3. RESULTS

Models are evaluated on the dataset provided by the MICCAI 2015 Gland Segmentation Challenge Contest [8]. As defined by the challenge authors, the dataset is split to 85 images in the training set (37 benign and 48 malignant sections) and 80 images in the test set, which is divided into two subsets. Set A contains 60 images (33 benign and 27 malignant sections), while set B contains 20 images (4 benign and 16 malignant sections). Fig. 1 shows a few samples from the dataset, with the corresponding labelling. A detailed description of the dataset can be found in [2]. Despite all efforts, we have not managed to reproduce results published in [9].

3.1. Impact of multi-task learning on single tasks

The original SA-FCN assumes uniform loss weights, i.e, \( \lambda = 0.5 \) for foreground and contour losses. To the best of our knowledge, no results describing loss weighting for this architecture were published. To investigate the impact of loss weighting on the performance of the model, the experiment has been carried out, in which we changed relative weighting rate between foreground and contour loss. Based on the results coming from the statistical significance test, it can be claimed that if there is a trend in the obtained results, it is very low. The noise of the optimisation, by far outweighs the choice of the loss weighting rate. More detailed analysis can be found in the paper [15]. The analysis leads us to the conclusion that the foreground and contour loss are very compatible with each other in many cases, which means that solving both tasks requires similar type of features.

Our analysis of SA-FCN architecture also included checking whether multi-task learning not only helps in the overall performance of the model, but also improves individual tasks performance. It has been found that on average, the model
Table 1. Classification (F1 Score) and segmentation (Dice Index) scores using three different post-processing methods. Scores are represented by mean value from five training followed by the standard error of the mean.

<table>
<thead>
<tr>
<th></th>
<th>Benign (A)</th>
<th>Malignant (B)</th>
<th>Benign (A)</th>
<th>Malignant (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA-FCN</td>
<td>0.760 ± 0.007</td>
<td>0.667 ± 0.023</td>
<td>0.836 ± 0.002</td>
<td>0.816 ± 0.008</td>
</tr>
<tr>
<td>DCAN</td>
<td>0.759 ± 0.006</td>
<td>0.687 ± 0.012</td>
<td>0.832 ± 0.003</td>
<td>0.814 ± 0.006</td>
</tr>
<tr>
<td>Our</td>
<td>0.860 ± 0.005</td>
<td>0.761 ± 0.007</td>
<td>0.851 ± 0.003</td>
<td>0.827 ± 0.004</td>
</tr>
</tbody>
</table>

with two decoders perform significantly better than the model with one decoder for both tasks (contour and foreground detection). Contour decoder helps the foreground decoder to ignore irrelevant parts of the image and put more attention to a particular region. Moreover, there is also visible assistance from the foreground side for the contour decoder, but not as strong as in the previous case.

3.2. Comparison of post-processing methods

This section presents, discuss and analyse the performance of the three different post-processing methods, 1) SA-FCN post-processing [9], 2) DCAN post processing [7], 3) our method.

3.2.1. Experimental Results

The results of the experiment are presented in Tab 1. Here, the mean scores from five training for the classification (F1 score) and segmentation (Dice index) tasks for three different post-processing methods is compared. Each column contains separate results for benign (dataset A) and malignant (dataset B) samples. Mean values are followed by the standard error of the mean.

Tab. 1 indicates that there is not much difference, in terms of performance, between the SA-FCN and the DCAN method. The most important result to emerge from the data is that our post-processing method perform significantly better on both tasks, regardless of the sample type. For the classification task, the difference reaches even 10% for the A dataset and 8% for the B dataset, which is a substantial increase. In case of Dice index, the difference is smaller, but still noticeable, 1.5% for benign samples and 1.3% for malignant samples.

3.2.2. Discussion

To better understand the results and to comprehend the practical differences between the applied post-processing methods, samples for which the change in the method had the greatest impact on the achieved result are presented in Fig. 4. The first row illustrates (from the left) the input sample, ground truth label and foreground prediction made by the SA-FCN. The next three rows represent the subsequent steps for three analysed post-processing methods. The first step of all three post-processing methods is fusion of contour and object probability maps. However, it is clearly visible that this operation causes the formation of many narrow and long objects along the main glands, meaning that all methods have to deal with a similar problem. This is due to misalignment issue, which is visualised in the Fig. 5.

From the Fig. 5, we can see that the contour prediction shape is very much in line with the ground truth. However, the contours prediction is slightly shifted in relation to the ground truth. The black pixels (which represent properly predicted pixels) are in the minority compared to the yellow-orange group of pixels (wrongly predicted). This may be due to down-sampling and up-sampling strategies used in the SA-FCN architecture. Max pooling, followed by the transpose convolution have the disadvantage that they are not able to ideally reproduce the location of the object, as consequence a small misalignment may appear. Both, the SA-FCN and DCAN methods turns out to be not sufficient enough to get rid of long narrow objects, which in turn are wrongly classified as separate instances of the glands.

The last row of Fig. 4 shows the subsequent steps in our post-processing method. After contour subtraction, very small objects are removed. The opening operation, meaning erosion followed by dilation, is the most crucial and allows to solve the aforementioned problem. As visible in the final labelling, the narrow long elements are successfully removed, which in turn limits the false positives and has a beneficial effect on the classification and segmentation scores, yielding better results than the DCAN and SA-FCN proposed post-processing methods.
3.3. Comparison of Mask R-CNN and SA-FCN performance

There is a tendency to design task-specific models for instance segmentation of gland images instead of utilising general purpose models, such as the Mask R-CNN, designed for natural images segmentation [3]. This is also one of the reasons why an idea for applying contour detection task has been introduced by data scientists, which is very characteristic for models designed explicitly for gland segmentation. The SA-FCN is one of such models. This experiment is devoted to compare performance of the Mask R-CNN and the SA-FCN architectures and verify whether Mask R-CNN can achieve a similar performance. The Mask R-CNN model has been adjusted to conduct instance segmentation on the gland dataset. Bayesian Optimisation has been used to perform hyper-parameters tuning. Random rotation, vertical and horizontal flipping has been performed as data augmentation. What is more, transfer learning has been used by using pre-trained weight on the COCO dataset. For the purpose of this experiment, our SA-FCN implementation (written in PyTorch) was used along with our post-processing method.

Both models were trained five times and the obtained results are averaged and presented in the Tab. 2. The Mask R-CNN architecture achieves on average higher results, than the SA-FCN. This is particularly evident in the case of classification task, where F1 score for the Mask R-CNN is higher on average than the SA-FCN by 2.7% in case of benign samples and 5.6% in case of malignant samples. On the other hand, the segmentation performance does not differ as much as the classification performance. The best proof for this is the fact that the Dice index score for the Mask R-CNN is higher than the SA-FCN by an average of 2.3% for dataset A, while for the dataset B, the SA-FCN obtained slightly better result, by 1.9%. Although, we were not able to reproduce the original results obtained by SA-FCN authors, our implementation achieves reasonable results. Interestingly, Mask R-CNN achieved better results than our SA-FCN implementation, but it still lacks around 3% to the state-of-the-art. What is important, the difference between the results of the Mask R-CNN and the results published in the article is so small that the performance of Mask R-CNN can be considered as comparable.

3.3.1. Discussion

In analysing the obtained results we pay special attention to the impact of post-processing. The Mask R-CNN does not require as much post-processing actions as the SA-FCN architecture. To better understand these observation we show on visualise a one sample (Fig. 6) before and after post-processing. Firstly, SA-FCN model has trouble with identifying large glands, instead separates them into several small ones which has a negative impact on the classification. Secondly, the SA-FCN prediction is more noisy and requires more post-processing effort to obtain satisfying results. The Mask R-CNN model generates instance segmentation predictions that are very similar to the final results. It requires only removal of overlapping elements or joining objects. The SA-FCN post-processing method consists of removing many small elements, filling the holes and what is the most important, model itself solves the task of semantic segmentation, not instance segmentation. Glands are divided into separate instance thanks to the post-processing actions. The

<table>
<thead>
<tr>
<th></th>
<th>F1 Score</th>
<th>Dice Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Benign(A)</td>
<td>Malignant(B)</td>
</tr>
<tr>
<td>M. R-CNN</td>
<td>0.888±0.003</td>
<td>0.817±0.002</td>
</tr>
<tr>
<td>SA-FCN</td>
<td>0.860±0.005</td>
<td>0.761±0.007</td>
</tr>
</tbody>
</table>

Table 2. Comparison of the Mask R-CNN and the SA-FCN models’ performance in terms of classification (F1 Score) and segmentation (Dice Index). Scores are represented by mean value from five training followed by the standard error of the mean.

Fig. 6. Visualisation of the same sample prediction before and after post-processing for the Mask R-CNN and SA-FCN models.
performed analysis leads to the conclusion that the SA-FCN performance strongly depends on the post-processing techniques, so a full disclosure of them is very important and necessary for a proper understanding of the principles of operation. Mask R-CNN seems to be more stable and reliable in its operation. Different architectures solve the same problem in various ways and we need to pay special attention to the appropriate way of comparing their results.

4. CONCLUSIONS

In this paper, we have outlined that post-processing can have a great impact on the final performance of the deep learning architecture, as it often hides many imperfections of the model. Our research showed that the selection of specific post-processing activities should be carefully adapted to the analysed dataset. When creating an algorithm, we suggest that specific cases are visualised and then experiment with different configurations. We find that it is misleading to perform a comparison of two DNN architectures, if one of them strongly depends on post-processing actions, without full disclosure of the raw output of the network. In addition, our work led us to conclude that domain specific models, in some cases, provide only marginal benefits at a cost of post-processing. Mask R-CNN, as a representative of a general segmentation model, obtains comparable results to current state-of-the-art. One advantage of Mask R-CNN is that the output requires little post-processing and can be easily adjusted and applied on any other dataset. Thus, it is worth considering to use and improve generic models such as Mask R-CNN as opposed to design complex architectures when tackling new domains.

5. REFERENCES


