MULTI-LOSS CONVOLUTIONAL NETWORKS FOR GLAND ANALYSIS IN MICROSCOPY

Aïcha BenTaieb, Jeremy Kawahara and Ghassan Hamarneh

Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Canada

ABSTRACT

Manual tissue diagnosis is the most prevalent approach to cancer diagnosis. However, it mainly relies on a subjective visual quantification of specific morphometric features, which often leads to a relatively limited reproducibility among experts. In most computational techniques proposed to automate the diagnostic procedure, accurate segmentation is paramount as a precursor to the extraction of relevant morphometric features. Since the ultimate goal of segmentation is generally classification, yet a given class imparts an expected tissue appearance beneficial to segmentation, we pose the problem of automatic tissue analysis as the joint task of segmentation and classification. We propose a novel multi-objective learning method that optimizes a single unified deep fully convolutional neural network with two distinct loss functions. We illustrate our reasoning on the task of colon adenocarcinomas diagnosis and show how glands' classification can facilitate their segmentation by adding class-specific spatial priors. The final classification also benefits from this joint learning framework yielding an improvement of 6% over classification-only models.

***Index Terms*—** Deep Learning, Histopathology, Classification, Segmentation.

1. INTRODUCTION

Almost all forms of cancer are diagnosed by the analysis of tissue biopsies. During diagnosis, a pathologist determines the presence of neoplasm in a tissue section on the basis of different criteria such as cellular abnormalities or proliferation. Further analysis of tumour-specific morphometric features are used to determine the tumour type or grade of differentiation. For example, the analysis of histology glands has proven to be a reliable bio-marker during the diagnosis of tumours developed in the glandular structures of epithelial tissues such as colon, breast or prostate carcinomas. Overall, pathologists' diagnosis of tumours involves a simultaneous identification of morphometric features and classification, which relies on a subjective visual-cognitive analysis of tissues. In addition to the potential for a false negative diagnosis, a direct shortcoming to this manual procedure is the limited intra- and inter-observer reproducibility, which has lead to the development of different automated diagnosis methods.

Generally, existing automatic methods for analyzing tumours from histopathology slides treat image segmentation and classification as two independent tasks. While segmentation is often done without the knowledge of the tumour type, classification usually relies on features extracted from pre-segmented images. For example, a large number of works [1, 2] focused on designing robust segmentation techniques for gland segmentation. Most of these works rely solely on combining pixel-level data with class-oblivious appearance priors to detect specific constituents of the tissue (e.g. glandular lumen, cellular nuclei, and stroma). However, as recently demonstrated by Sirinukunwattana et al. [2], most of these existing gland segmentation techniques fail when applied to different tumour grades as the tissue structures shape regularity assumption no longer holds for pathological cases. The recent success of machine learning techniques, which tightly integrate feature learning and classification in a single framework, has allowed for bypassing the pre-segmentation and feature design steps, i.e. the learnt discriminatory features are derived directly from the input image. A key example is the use of convolutional neural networks that have been successful in the pattern recognition task of classifying mitotic cells, and have also been extended to segmentation being modelled as a pixel-level classification task [3, 4]. To the best of our knowledge, none of these deep learning-based models attempts to integrate classification priors in the segmentation in an end-to-end framework.

While it is generally accepted that features based on segmentation are critical for classification, the inverse, i.e. classification can benefit the segmentation, is much less explored. A given tissue class imparts an expected class-specific tissue appearance beneficial for guiding the segmentation. This dilemma: classification requires segmentation-based features but accurate segmentation requires knowledge of the class to-be-segmented, justifies a joint segmentation-classification approach. In contrast to existing works, we propose a joint deep learning model where the segmentation and detection of morphometric features are coupled with tissue classification. We make the assumption that there exists certain candidate segmentation regions that benefit the classification of the underlying tissue and, in reverse, the knowledge of the tumour type allows a finer automatic tuning of a class-dependent segmen-

We thank NSERC for funding.

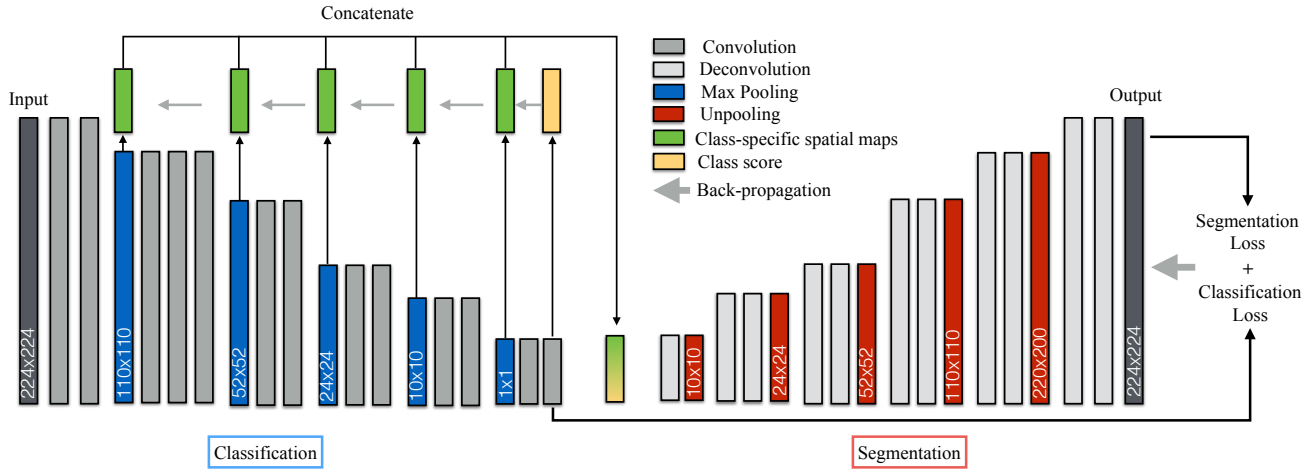

Fig. 1. Multi-loss network architecture.

tation by reducing the segmentation search space.

In this work, we focus on the analysis of glands from colon adenocarcinoma slides. We propose an end-to-end framework based on a deep fully convolutional network for colon adenocarcinoma segmentation and classification. Given an input tissue slide, our system predicts a tumour type (benign vs malignant) and provides a segmentation of the relevant glandular structures. This paper shows how the joint learning of a segmentation and classification can be modeled effectively in a unified framework using a novel deep learning architecture and a multi-loss objective function.

2. METHOD

Figure 1 presents the overall architecture of our network with a classification and a segmentation components organized symmetrically. The classification component, composed of layers of convolution and subsampling, identifies the tumour type. Then, the segmentation component performs the reverse operations with layers of deconvolution and upsampling to produce pixel-wise labelling of the identified glands. We define cross-network spatial activation maps to integrate the classification feature maps into the segmentation.

Given an input RGB image x, our goal is to simultaneously predict a class label $\hat{y} \in \{0,1\}^2$ ($[1,0]$ for benign $[0,1]$ for malignant) and a 2-channel segmentation mask $\hat{S}$ where each channel corresponds to a label map for gland and background. We use a multi-loss objective function to train the network and optimize both segmentation and classification in an end-to-end (i.e. across the whole network) fashion. The next sections describes the detailed configurations of the proposed network and its components.

2.1. Classification

To effectively train our full network, reduce training time and avoid vanishing gradients as observed in very deep neural networks, we pre-train the classification component. Given an input image x and a ground truth class label y, the classification model outputs a normalized score vector $Q(x) \in \{0,1\}^2$. The classification's objective is to minimize the error $\mathcal{L}_c$ between ground truth y and predicted score $Q(x)$ using the logistic loss:

$$\mathcal{L}_c(x_i, y_i) = -y_i \log\left(Q(x_i)\right) \quad (1)$$

$$Q(x) = \frac{\exp\left(z_c^y(x)\right)}{\sum_{l=1}^{L} \exp\left(z_c^l(x)\right)} \quad \text{for } y \in \{1,2\}, \quad (2)$$

where z_c^y is the classification component's output activation for label y.

2.2. Cross-network feature maps

We exploit the pre-trained classification component to extract class-specific spatial maps (i.e. locations of discriminatory regions) and inject these cues to the segmentation component. The stacked layers of convolution and subsampling learned by the classifier preserve the spatial configuration of glands' locations. However, the layers' activation maps may contain mixed activations from all class labels in the image, so we need to identify which activations are relevant for the given image. To this end, we define *class-specific spatial priors* which correspond to the averaged output activation maps from all pooling layers of the classification component. We combine these pooling layers such that they include only information relevant to the class of the input image. More concretely, given an image x, the pre-trained classifier outputs a normalized score vector $Q(x)$. Our goal is to rank pixels in x based on their contribution to the final score $Q(x)$. As shown by Simonyan et al. [5], determining the importance of each pixel of the input image x on the output classification score $Q(x)$ corresponds to computing the class-score derivative with respect to the image.

We extract class-specific spatial priors, f_k^l for a given class l from each pooling layer k of the classifier by computing the derivative of the class score $Q(x)$ with respect to the

activation value z_k of pooling layer k:

$$f_k^l = \frac{\partial Q(x)}{\partial z_k^l(x)}, \tag{3}$$

This operation can be performed using back-propagation from the final class score layer to each pooling layer. Intuitively, computing the derivative operation in eq. 3 amounts to measuring the contribution of the activations, in each pooling layer, to the final class score.

We use this class-specific spatial information as a prior to the segmentation component by upsampling and average pooling all activation maps $\{f_1^l, \ldots, f_k^l, \ldots, f_K^l\}$ obtained from the classifier's pooling layers (see Figure 1). We will refer to the final spatial prior as f.

2.3. Segmentation

The segmentation component takes as input three elements: i) the input color image x of size $W \times H \times 3$; ii) the class-specific spatial prior f for the given image; and iii) the normalized class scores output from the classifier. This component predicts a 2-channel label map $P(x) = \{P_g(x), P_b(x)\}$ of size $W \times H$ for gland and background classes where $P_g, P_b \in \mathbb{R}^2$.

The provided ground truth segmentation (background vs. gland) does not directly encode the specific appearance characterizing the glands, i.e. the gland comprises a central area corresponding to the lumen surrounded by epithelial cells. However, we incorporate this gland appearance prior using the class-specific spatial prior (obtained in Section 2.2). We hypothesize that the class-specific activation maps will enable us to encode glands' appearance by showing stronger activations for pixels located at the center of detected glands. Thus, we weight the ground truth segmentation S by the class-specific activations f and define the following loss function:

$$\begin{aligned}\mathcal{L}_s(x_i, S_i, f_i) = -\sum_{j=1}^{\Omega} (S_{ij} \times f_{ij}) \log(P_g(x_{ij})) \\ +(1 - S_{ij} \times f_{ij}) \log(P_b(x_{ij})),\end{aligned} \tag{4}$$

where Ω is the total number of pixels in the input image; P_b and P_g correspond to probabilities of a pixel belonging to the background b, or to the lumen g of the gland respectively.

2.4. Joint training and prediction

After pre-training the classification layers and extracting the cross-network activation maps using back-propagation as defined in Section 2.2, we train the full network including the segmentation layers. We use a weighted multi-loss objective function which jointly optimizes eq. 1 for classification and eq. 4 for segmentation, as follows:

$$\mathcal{L}(x, y, S) = \lambda \mathcal{L}_c(x, y) + (1 - \lambda) \mathcal{L}_s(x, S, f), \tag{5}$$

where λ is a user-specified coefficient used to weight the relative importance of each loss in the multi-objective function.

At test time, we compute the class-specific spatial activation maps and obtain segmentation maps identifying malignant and benign glands. The final predicted segmentation mask $\hat{S}$ is obtained by identifying the maximum score in each pixel out of the 2-channel scoring map $P(x)$:

$$\hat{S}_j = \arg\max \{P_g(x_j), P_b(x_j)\} \quad \forall j \in \Omega. \tag{6}$$

3. EXPERIMENTS

3.1. Dataset

We evaluate our method on the available Warwick-QU dataset [2] which consists of 37 benign and 48 malignant H&E stained colon adenocarcinomas. Each slide was scanned at 20x microscope magnification and annotated by an expert pathologist who graded the tumour type and handmarked the ground truth segmentation. Each slide was approximately 500×700 pixels. After resizing each image to 500×500 pixels, we extract non-overlapping image crops of size 250×250 from each slide. We split the full dataset into train, validation and test sets with the ratios 70,10, and 20%.

3.2. Implementation

Images and their corresponding tumour label and segmentation are used to train the network. To include a sufficient amount of variability in the training set samples and gain robustness during training, we augment the training set using a series of spatial (affine and elastic) transformations and colour perturbations. We use Caffe library to implement and train our network with stochastic gradient descent with momentum as solver. Input images are cropped to 224×224 to fit our GPU memory constraints and a batch size of 1 is used. Accordingly, we set the momentum to 0.99 in order to consider a large number of training samples during the gradient update. We set λ in eq. 6 to 0.5. The network converges after approximately 20,000 iterations of the solver and training takes 5 days on our single 4 GB-memory GPU.

3.3. Evaluation

We evaluate the contribution of our multi-loss network on 1) the tumour classification accuracy and on 2) the segmentation accuracy for malignant and benign glands at both: 2a) the pixel-level, i.e. pixel label classification and 2b) the object-level via the Dice similarity coefficient (DSC). For a fair comparison, we use as baseline convolution-based networks recently proposed for image classification (AlexNet [7]) and biomedical image segmentation (UNet [4]). We also evaluate the performance of each of our network's components individually as classifier or segmentation network only. For

Networks	Class ACC (%)	Pixel ACC (%)	Benign DSC	Malignant DSC
AlexNet [7]	83.0	-	-	-
Multi-Loss-Class	83.0	-	-	-
UNet [4]	-	78.0	0.62	0.55
Multi-Loss-Seg	-	86.8	0.70	0.56
Multi-Loss-Joint	**89.0**	**92.0**	**0.90**	**0.76**

Table 1. Segmentation and classification performance. Multi-Loss-Seg and Multi-Loss-Class correspond to our model segmentation and classification component only wherease Multi-Loss-Joint refers to the proposed joint learning model.

these experiments, we train the network using the classification or the segmentation loss individually for each task. In all our experiments, each network's parameters were tuned on the validation set and Table 1 reports the best performance achieved after convergence.

3.4. Results

Our results confirm the advantage of jointly learning classification and segmentation networks. We show better performance (up to 6% increase in accuracy and 6 to 20% increase in segmentation) when using our multi-loss implementation for both classification and segmentation tasks. We also note that our classification and segmentation component used individually show competitive results with baseline networks proposed for image classification and segmentation. We observe a clear gain in DSC accuracy when segmenting malignant glands, which are often harder to segment due to their complex shapes [2]. We believe that the supervision added by the class-specific spatial activation maps allows to encode glands shape priors that contribute to the segmentation.

Figure 2 presents qualitative results of the segmentation outputs generated for benign and malignant glands. We also show the class-specific spatial priors generated from the trained classifier component. The class-specific spatial priors show the location of detected glands. This interesting property allow us to reduce the search space of candidate segmentations for a given image by reducing variations of input distributions and detecting relevant objects.

4. CONCLUSION

We proposed a multi-loss convolutional network for joint classification and segmentation of colon adenocarcinoma glands. By including class-specific spatial priors, we were able to train more effectively our segmentation network and restrict the set of candidate segmentations based on their discriminative ability. We showed how classification and segmentation can be learned simultaneously and resulted in improved performance for both tasks. We believe our system can benefit different histopathology image analysis tasks

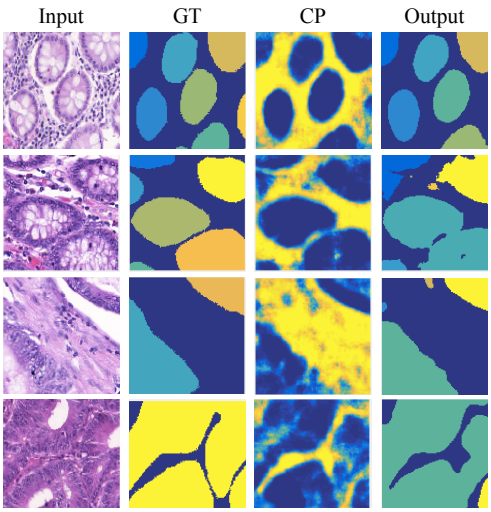

Fig. 2. Qualitative segmentation results. GT are the ground truth segmentation, CP are the class-specific spatial priors.

involving tumour diagnosis and morphometric features identification. To train our system, we rely on strongly annotated datasets provided with class labels and binary segmentation which limits the applicability of our work to weakly-labeled datasets but raises new challenges for future works.

5. REFERENCES

[1] Cigdem Gunduz-Demir et al., "Automatic segmentation of colon glands using object-graphs," *MIA*, vol. 14, no. 1, pp. 1–12, 2010.

[2] Korsuk Sirinukunwattana, David Snead, and Nasir Rajpoot, "A stochastic polygons model for glandular structures in colon histology images," *IEEE TMI*, vol. 34, no. 11, pp. 2366–2378, 2015.

[3] Dan C Cireşan et al., "Mitosis detection in breast cancer histology images with deep neural networks," in *MICCAI*, pp. 411–418. 2013.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint: 1505.04597*, 2015.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint: 1312.6034*, 2013.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.