Université d'Ottawa | University of Ottawa

DEEP LEARNING FOR MELANOMA DIAGNOSTICS

A.Phillips[1], I. Teo[2], J. Lang[1]
[1]School of Electrical Engineering and Computer Science, University of Ottawa [2]Department of Pathology and Laboratory Medicine, University of Ottawa

uOttawa

OVERVIEW & CONCLUSIONS

RESEARCH QUESTIONS

Is it possible to accurately predict segmentation masks for multiple tissue types whose structures are vastly different in scale, and morphology with dramatic variations even between tissues of the same type as is the case in the pathology of the skin (see figure 1.)?

While we do not know the segmentation accuracy required for prognostic measurements, can we produce segmentation masks that are qualitatively, sufficiently accurate to perform measurements such as Breslow thickness, and Clark level for melanoma?

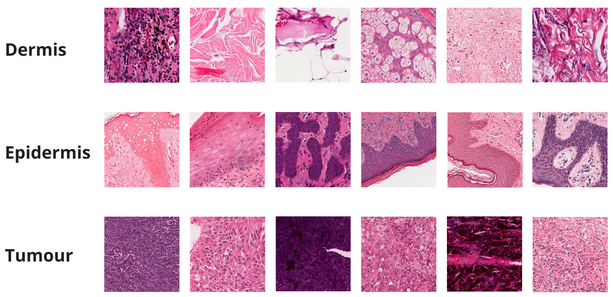

Figure 1. Plate showing the variations within and between multiple tissue types.

METHODOLOGY

1. Curated a dataset of tissue specimens containing cutaneous melanoma, including full resolution annotations of tumour, epidermis, and dermis tissues by an expert pathologist.
2. Developed a neural network architecture specifically designed to incorporate features at multiple levels of granularity present in the images contained in our dataset for supervised learning.
3. Trained a model using our dataset and architecture to detect and localize structures important for melanoma diagnostics.
4. Assembled a panel of four pathologists to independently measure Breslow thickness on four cases from the test set, and computed standard segmentation metrics on model output.

CONCLUSIONS

We have demonstrated a method capable of detecting and localizing melanoma tumours and other tissue structures important for prognostic measurements using a custom fully convolutional network on a dataset that we have curated.

Given the qualitative and quantitative results it is clearly possible to overcome the discriminative challenges of the skin and tumour anatomy for segmentation using modern machine learning techniques. Further we have shown it is possible to approach a level of accuracy to manually perform the Breslow thickness measurement. More work is required to improve the network's performance on dermis segmentation.

This work is a starting point for potential future applications in cancer pre-screening or within QA/QC workflows for prospective or retrospective review of past cases for diagnostic discrepancies.

30 SECOND VERSION

- We curated a dataset of whole slide images (WSI) of melanoma excisions
- We trained a custom fully convolutional network (FCN) using WSI patches
- Our model accurately predicts tumour and epidermis locations with sufficient accuracy for measuring the Breslow thickness prognostic metric

REFERENCES

1. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3431, 2015

THANKS TO BRESLOW PANEL: Y. Ayroud, B. Burns, and S. Petkiewicz

CONTACT: aphil037@uottawa.ca

SCAN TO DOWNLOAD THIS POSTER.

BY-NC-SA

RESULTS & DISCUSSION

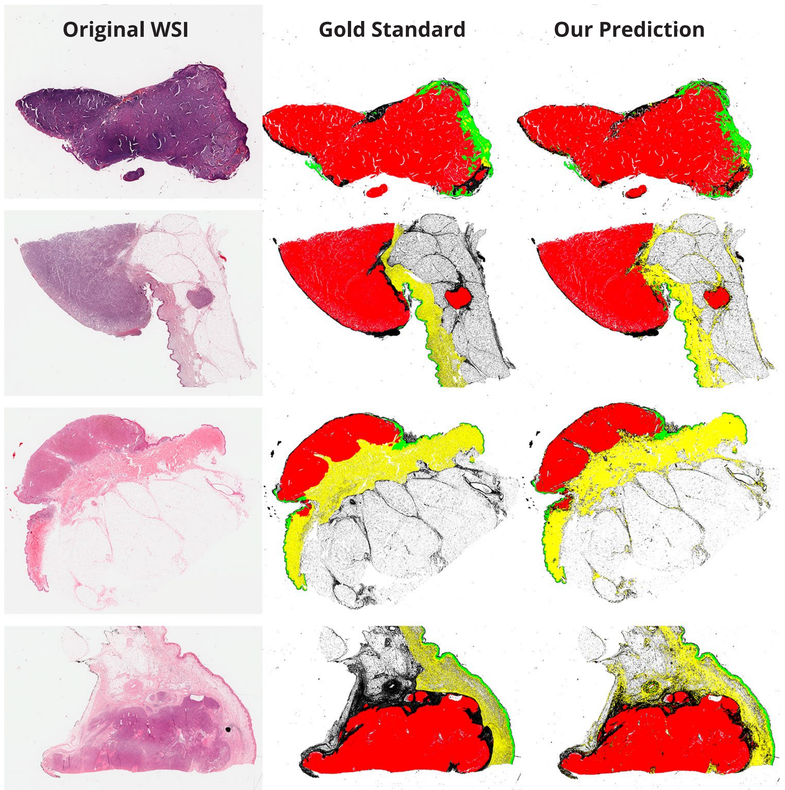

Figure 2. Segmentation results from our model compared with the pathologist's gold standard annotations Tumour shown in red, epidermis in green, dermis in yellow, other structures in black.

Qualitatively, figure 2 shows a strong correspondence between the gold standard and the predictions. We observe that the model is able to differentiate between tissue types and between highly varied morphology between tissues. However the model struggles to infer the true depth of the dermis. This is related to the presence of "dermis cells" in deeper layers of the skin.

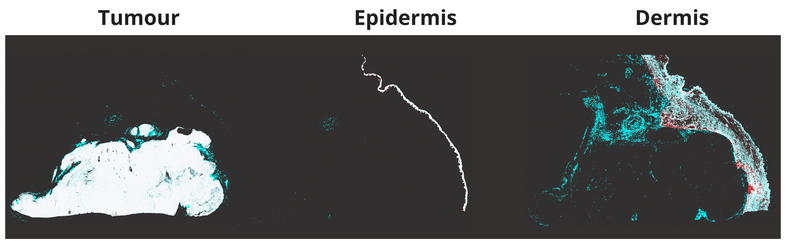

Figure 3. Visualization of segmentation errors by pixel. White represents true positive, teal represents false positive and red represents false negative by respective tissue type

Errors in tumour and epidermis tissue are largely found around boundary regions which do not affect practical accuracy (see figure 3). Less common serious errors could be addressed by increasing the size of the dataset, tuning the FCN and improving normalization.

The results of a panel of four pathologists measuring Breslow on the output were diagnostically equivalent. When removing cases where measurements were not appropriate, IR-Reliability was 87.5%

NETWORK COMPARISON

Our network design outperformed our comparators based on standard metrics for semantic segmentation. The network has greater discriminatory power with improved performance while requiring less computing power (see figures 4 and 5).

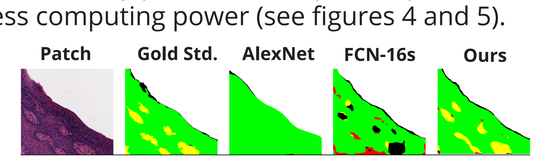

Figure 4. Example discriminatory differences at the patch level. Note the yellow dermal regions.

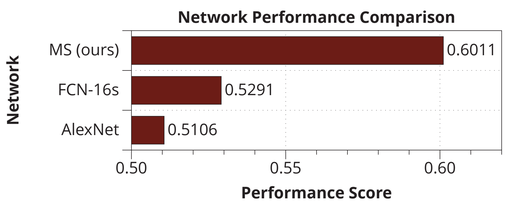

Figure 5. Zoomed net performance where score is mean pixel accuracy over mean intersection over union.

METHODS

DATASET

- From over 1000 WSIs obtained from The Cancer Genome Atlas (TCGA), we assembled 50 cases from 49 individuals that were subsequently annotated by apathologist using a Photoshop based workflow (faster and easier to use than specialized tools).
- WSIs were divided into training, validation and test sets at a ratio of 70:15:15 percent respectively.
- To resolve the large class imbalance present prior to training, we implemented a method of oversampling the minority class (epidermis) and undersampling the most common classes (fig. 6).

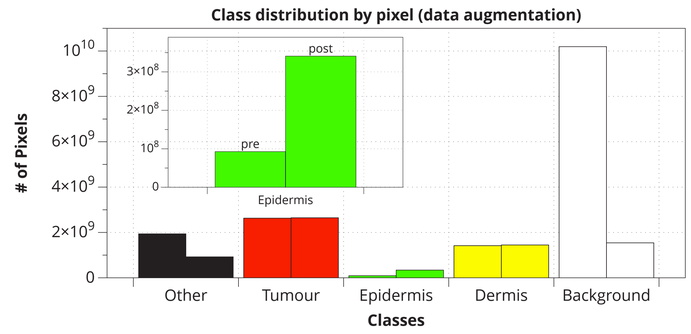

Figure 6. Class distribution before (left bar) and after augmentation. Inset highlights the epidermis class.

FULLY CONVOLUTIONAL NETWORK

- We designed a new FCN architecture to include features at multiple levels of granularity via deconvolution using multiple pixel strides (32,16, 8). We then combined these scores using a weighted, element-wise summation, where per-stride weights are hyper-parameters for our model (see figure 7).
- We compared the performance of our design against two common architectures used for similar tasks both in general computer vision and medical tasks to validate our design for use in melanoma segmentation.

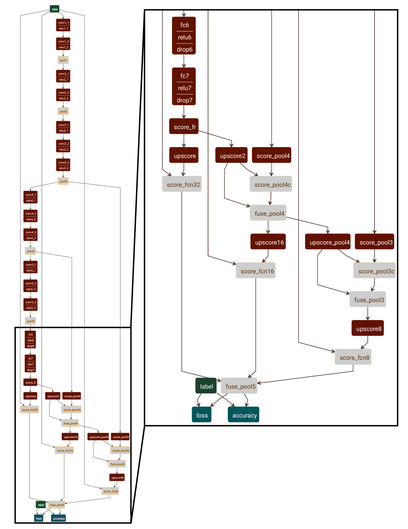

Figure 7. FCN showing multi-stride fusion.

TRAINING AND INFERENCE

- The WSIs and the corresponding annotations were divided into patches of size 512x512 to facilitate ingestion into the network.
- Using the patches from the training set, a model was trained via stochastic gradient descent with dropout over a total of 40 epochs
- Network weights were initialized with publicly available values trained on the ImageNet and PascalVOC datasets for the base convolutional layers that our network shares with FCN-16 [1].
- At inference time, WSIs from the test set are "patchified" and sent to the trained model one-by-one. The model returns a segmentation mask for each patch, which are then assembled into the resulting whole slide segmentation mask (see figure 8).

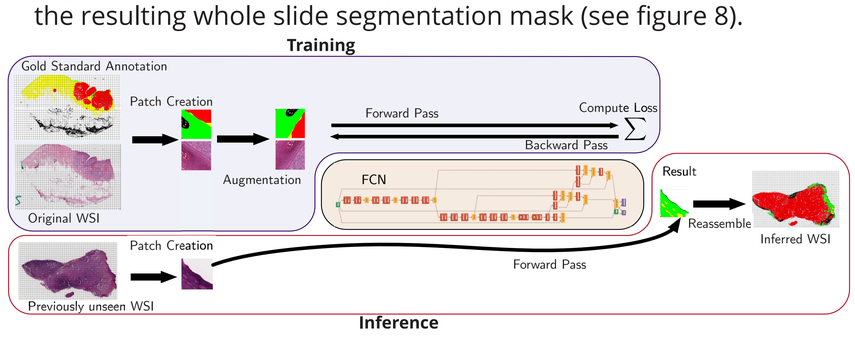

Figure 8. Visualization of the patch flow through the network at the training and inference stages.