Self-Supervised Semantic Segmentation of Lymph Node Metastasis

Sashaank Pasumarthi and Tanmay Jaiswal
Department of Computer Science, Columbia University

I. ABSTRACT

Each year, there are more than 200,000 cases of breast cancer identified in the United States. One of the core decisions that affect the course of treatment for breast cancer is whether or not the cancer has metastasized to other regions of the body. However, arriving upon this decision is quite arduous and often error-prone. To assist physicians in making better diagnostic evaluations, Liu et al. proposed a convolutional neural network-based architecture to automate the diagnostic workflow and provide a second opinion in *Detecting Cancer Metastasis in Gigapixel Pathology Images*[1]. Achieving an image-level AUC score above 97%, the paper established a new state-of-the-art for cancer metastasis detection. However, achieving these results required plentiful annotated data and the approach focused on localization and classification rather than pixel-level segmentation. In our work, we try to address these two issues by proposing a method for self-supervised semantic segmentation of lymph node metastasis. We believe our method can be generalized to most medical imaging classification tasks that suffer from a limited supervised training set. On the lymph node metastasis task, we show that we can achieve a pixel-wise F1 score of 69.94% while using just 3% of the annotated training data.

II. INTRODUCTION

As mentioned earlier, the treatment procedure for breast cancer heavily relies on the level of metastasis. Generally, detecting cancer metastasis is a highly challenging and time-consuming task that requires specialist pathologist knowledge. The procedure involves pathologists microscopically observing many slides of lymph node tissue adjacent to the breast to identify any spread of the tumor cells. This involves pathologists carefully sliding through each slide and observing it at different magnification levels to assess metastasis.

Recently, deep convolutional neural networks have shown impressive performance on many computer vision tasks including semantic segmentation and object detection. As such, many applications of deep learning in medical imaging have come to the forefront to alleviate the burden of clinicians and improve medical diagnostic tasks. While these models have generally performed well, one of the biggest limitations to further integrating deep learning into the medical field is the lack of annotated training data. While there are ample medical images available for training, most of this data is unlabeled. This is simply because labeling such data requires doctors or lab technicians with highly specialized knowledge and adequate time. Unfortunately, this is a quite expensive endeavor which severely limits the extensive impact deep learning could have on the field, as all models need sufficient training data to be deployed.

In this particular case, we attempt to use a self-supervised approach to overcome the lack of annotated training data for detecting breast cancer metastasis.

Before jumping into our procedure and evaluation, it is important to understand the purpose and goal of all this automation applied to medical tasks. The goal of machine learning is not to replace

physicians, but rather to aid and assist physicians in making the proper diagnosis. As such, the recommendation for integrating these tools into the diagnostic workflow is not as a primary diagnosis, but rather as an automated second opinion. The idea is that physicians should be making a judgment completely independent of the diagnostic tool, so as to prevent overreliance on the technology and keep medical judgment free from influence. By instilling this tool into the pipeline as a secondary opinion, it provides a cheap way to reaffirm the initial diagnosis or take a closer look at the previous outcome.

III. METHODS

1. DATASETS

The dataset used in this study was the CAMELYON16 dataset from the lesion-level tumor detection task. The dataset consists of gigapixel pathology images of size 100,000 x 100,000 pixels. The dataset contains 270 tumor slides with pixel-level tumor annotations, although most of the annotated masks were not accessible for our use.

2. Preprocessing Method

In the first part of our method, we use a similar pre-processing technique to Liu et al [1] in order to load our data into the models. The only difference lies in the amount of training data used for the task. While the paper is able to use a majority of the 270 annotated slides for training, commodity hardware meant that we were relegated to using a much smaller subset of 10 annotated slides for training. This is because each gigapixel pathology slide takes between 1-3 GB of RAM, and using the high-RAM option with Google Colab provides for a maximum of 25 GB.

Because of the large size of the slides, rather than trying to feed the entire slide into the neural network, 300-pixel x 300-pixel patches are extracted from each slide and aggregated to create the training set. When we sample patches from the training slide in order to create the training dataset, we use a random sampling approach. This is so we can create a decent representation of the larger slide without having to analyze every part. First, we select either “tumor” or “non-tumor” with equal probability. Then, we randomly sample indexes in the slide until we are able to extract enough patches for each category. The reason for sampling for each class separately is to address the tumor imbalance in the slides. For most slides, there are many more non-tumor regions than tumor regions and we want to ensure our model receives a representative sample.

In order to further bolster our training set, we also use data augmentation to increase the prevalence of tumor patches. For every sampled patch, a set of 8 standard data augmentations are performed in order to 8x the size of the training data. These include rotating the input patch by all multiples of 90 degrees, applying a left-right flip, and another set of 90-degree rotations to create 8 augmented permutations of every training patch. Then, a series of color transformations including random brightness, contrast, saturation, and hue are performed before adding the patches to the overall training set.

The above method is used to extract both our supervised and unsupervised training sets. For the supervised training set, we also extract the pixel-annotated tumor mask corresponding to each patch.

3. Simple Image Reconstruction

The next few steps in our approach take inspiration from *Unsupervised Pre-training for Fully Convolutional Neural Networks* [3]. The first part of our task is to train a UNet convolutional neural network for image reconstruction. Essentially, our large unsupervised training corpus is fed into the UNet architecture with the task of reconstructing the image just as it appears. UNet is used for many image segmentation tasks and was specifically

introduced by Ronneberger et al [2] for biomedical imaging. While our approach pretty much uses the standard UNet architecture, as shown in Figure 1, one slight modification we made was to omit the last skip connection to prevent the model from simply copying over the features and only using those for the reconstruction task.

Justification - The UNet is a standard encoder-decoder model. An encoder block performs a series of convolutions on the input and then downsamples it. The decoder block performs convolutions on the input and either up samples it uses transposed convolution to increase the output size. The intuition is that as we go deeper, we learn higher-level features. These higher-level features are combined with larger size outputs from previous encoder blocks to restore the spatial information lost due to the downsampling. According to this understanding, we assume that the features after the first convolution block are sufficient to represent the entire image and even learn higher-level features from. In a sense, the output of the first convolution is a good representation of the input image. It contains both spatial information as well as few deeper features. Otherwise, it would not be possible to perform segmentation with high accuracy. If this assumption holds, during image reconstruction pretraining, the model can ignore the deeper features and use only the skip connection to reconstruct the output.

We found that this was true and the weights of the deeper convolution blocks were near zero. The gradients would not propagate through the depth of the network. Removing the skip connection forced the model to reconstruct the output using the deeper features resulting in non-zero weights and proper learning.

4. Fine-Tuning

After training a base UNet model on the image reconstruction task, we use these pre-trained weights to initialize our UNet model for fine-tuning. We fine-tune the model on both the semantic

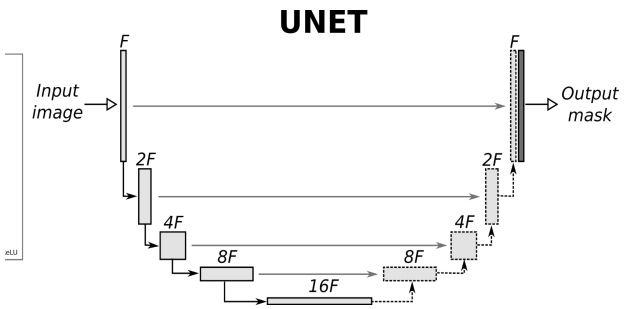

Fig 1. The standard UNet architecture

segmentation and image reconstruction task. While our true objective is to fine-tune for semantic segmentation to receive pixel-wise segmentation of tumor metastasis, we observed that our model tends to destroy previously learned information unless it is also fine-tuned on the image reconstruction task in parallel. This practice is also followed by Wiehman et al[3].

The input in the fine-tuning step is the limited supervised training data, which includes the pixel-annotated tumor masks. For our approach, we experimented with various sizes of our supervised training set, which starts at 3% of our overall training data and reduces.

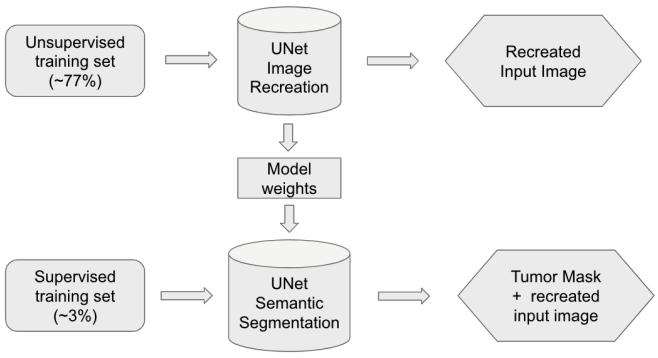

Figure 2: Overview of Training Method

IV. EXPERIMENTS AND EVALUATION

The objective of our project is to limit the amount of labeled training data required to achieve high model performance on tumor metastasis detection. As such, we experimented with various

supervised training set sizes to determine how much annotated data was truly required to create a high-performance model.

While we used the same base model trained for the image reconstruction task, we used different supervised training set sizes for the fine-tuning step and recorded the model performance with each. In addition, we used 5-fold cross-validation on the supervised training set of each size to average out performance.

For evaluation, we decided to quantify our model performance using pixel-wise precision, recall, and F1 score. Because of the large class imbalance between tumor and non-tumor pixels, accuracy tends to be a misleading metric.

Also, it made intuitive sense to prioritize recall because of the significant clinical implications. If the model has many false negatives, it could have lethal consequences due to misdiagnosis and lack of treatment.

We show the model's performance for different training set sizes in Figure 3 below.

	Without pretraining			With pretraining		
Supervised set size	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
200	57.71	88.76	69.94	59.37	89.83	71.49
120	54.65	81.0	65.26	55.35	84.79	66.981
80	49.87	85.63	63.18	56.56	87.22	68.62
40	41.12	76.8	53.58	46.71	77.06	58.16

Figure 3: Model Results

We compare the performance of our method against training a UNet model directly on the limited-size supervised training set. We chose this as the baseline because it applies the real-world constraints of limited annotated training data. We can see an increase in performance across the board. For every metric and training size, our approach has a substantial improvement over the baseline model.

In terms of annotated training set size, Figure 3 clearly illustrates the general pattern that the more annotated training data, the better the performance of the model. However, there is an exception as a set size of 80 seems to outperform the set size of 120. It is not immediately evident why that is the case. This could be the result of data imbalance since the supervised set was chosen at random and the probability of data imbalance increases with a decrease in sample size.

In order to better visualize the performance of our model, we can observe its output on both the image reconstruction and segmentation tasks.

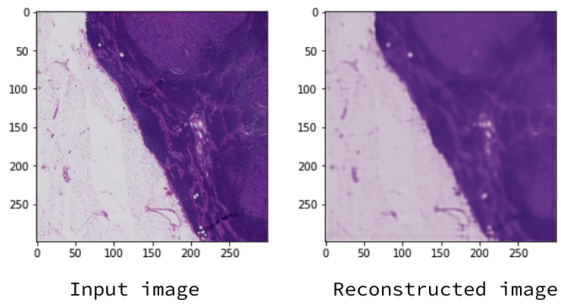

Figure 4: Image Reconstruction Task

Clearly, the model performs exceptionally well at the image reconstruction task. While it is not the exact image, this is because of the omitted skip connections which purposely added some noise to the model during pre-training.

The model is also quite proficient in the semantic segmentation task. For the most part, the model is able to isolate the regions which have tumor prevalence and provide fine-grain classification.

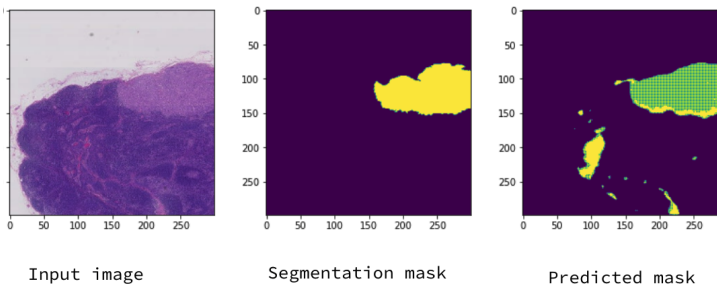

Figure 5: Semantic Segmentation Task

V. CONCLUSION

In this study, we combined the work of Liu et al [1] and Wiehman et al [3] to propose a framework for semantic segmentation to detect breast cancer metastasis. We were able to address the real-world constraint of limited annotated medical training images and showed that a self-supervised approach can be used to circumvent it to an extent. While we were only able to demonstrate this for the breast cancer metastasis task, we believe this method can generalize to other medical imaging tasks and further work is necessary to evaluate further.

In addition, we demonstrated how much even a few extra annotated training examples can boost performance noticeably.

VI. REFERENCES

1. Liu, Y. et al. Detecting cancer metastases on gigapixel pathology images. Preprint at https://arxiv.org/abs/1703.02442 (2017).

2. Ronneberger O., Fischer P., Brox T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

3. S. Wiehman, S. Kroon and H. de Villiers, "Unsupervised pre-training for fully convolutional neural networks", *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pp. 1-6, Nov 2016.