Classifying Histopathological images with Deep learning

Mohammad Mohaiminul Islam

1 Introduction

Depending on the imaging modality, organs and tissues are revealed structurally or functionally in the images produced in the medical domain. This data can then be used to guide clinicians with clinical evaluation, treatment, planning, and follow-up [1]. It also opens up the possibility of a quantitative study of this data which can contribute to a better analysis of structures and functions in normal and abnormal cases while avoiding intra/inter-expert variability. The liver parenchyma of patients with liver fibrosis is replaced with scar tissue [2]. Among the main causes of liver fibrosis are excessive alcohol consumption, severe steatosis and steatohepatitis, and viral hepatitis [3]. In severe instances of liver fibrosis, such as cirrhosis, there is a risk of hepatocellular carcinoma or liver failure. Thus it is very important to diagnose and assess liver fibrosis/cirrhosis using histopathology. A challenge, however, is that histopathological images are characterized by an inhomogeneous structure and a high degree of complexity. Computer-assisted medical image analysis aims at optimizing the use of medical data to solve these kinds of problems. Recent years have seen a significant increase in computer-aided diagnostic (CAD) systems that heavily rely on deep learning technologies to assist pathologists with analyzing these images.

2 Tools & Libraries

The developed models and related experiments have been implemented primarily using Python as the programming language, and PyTorch [4] as the deep learning framework. While U-net [5] with VGG [6] backbone was implemented from scratch ,U-net with ResNet [7] backbone were used from segmentation-models package [8]. Besides, other supporting packages like OpenCV, SciKit-image, matplotlib, etc., were used for miscellaneous work.

3 Dataset

The provided data set contained training, validation, and test set. First, three images were moved from the validation set to the test set, which will be referred to as extra-test data. The idea here is to have some unseen data to the model with a mask (ground-truth) so that the model's prediction could be understood and evaluated better. Next, the original images of dimension 1024*512*3 were processed into manageable patches of 256*256*3 to reduce the computational load and increase model performance. Further, this also helps alleviate the problem of a small training set. One observation during training was that the models were suffering from a large generalization gap. Data augmentation was introduced during training to combat the issue. Following data augmentation were applied: random rotations (between -20 degrees to +20 degrees along the three different orthogonal axes of the image where the randomness and selection of axes follow a normal distribution), vertical and horizontal

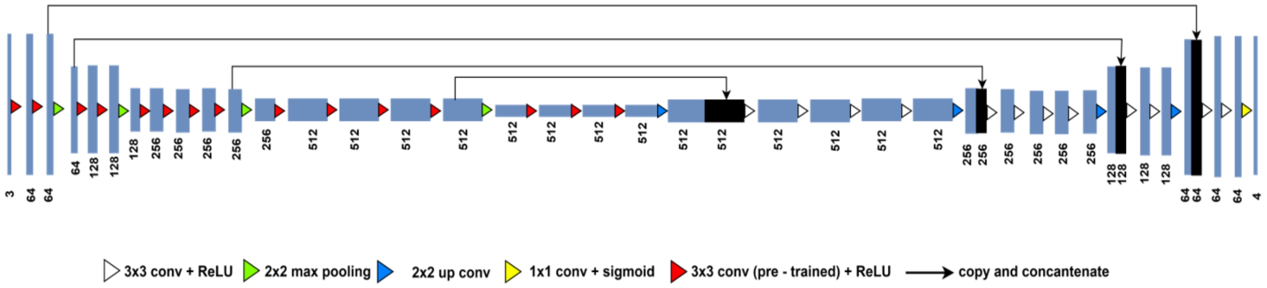

Figure 1: U-Net with VGG19 pretrained on ImageNet

flips, contrast adjustment (defined as $c = \mu + \alpha \times (v - \mu)$ where μ is the mean voxel value and α a scaling factor between 0.5 and 1.5).

4 Methods

The problem of classifying each pixel into one of the three types (Parenchyma, Fibrosis, Background) of tissues could be solved by building a pixel-wise classification model where we would have certain probabilities for every pixel belonging to each class. However, the pixels were divided into four classes in this study: class 0 – Unknown tissue, class 1- Parenchyma, class 2-Fibrosis, and Class 3-Background. This problem can also be viewed as a segmentation task since these pixels would ultimately make up a region in the image. U-net has been popular in general semantic segmentation and medical image segmentation/pixel-wise classification tasks. Researchers have modified the encoder of the U-net architecture with various well known models such as ResNet, DenseNet [9], InceptionNet [10] pretrained on ImageNet [11] to boost the performance. More advanced techniques such as Attention Network [12], Transformer based approach [13] and deep supervision [14] are being employed to improve the performance further. However, for the sake of simplicity, lack of enough training data, and time constraints, a more simplistic U-net was used here.

4.1 Experimental Setup

As final architecture U-net with VGG19 and ResNet34 were used. Fig 2 shows the detailed architecture of U-Net with the VGG19 encoder. Here the figure of U-Net with resnet34 is omitted due to the space limitation. For training the model, standard pixel-wise cross-entropy loss and a hybrid loss composed of pixel-wise cross-entropy and dice loss were used. The latter shows to improve performance slightly. The dice coefficient explains the level of overlap between our network prediction result and the ground truth label, which is one of our main optimization objectives. Cross-entropy loss allows the loss function to converge by maintaining a smooth gradient for all pixels. As a result, combining these two losses may fully use their benefits, such as handling imbalanced categories, smoothing gradients, optimizing segmentation details, and achieving improved network performance. Hence used hybrid loss is defined as follows:

$$\mathcal{L}_{hybrid} = \mathcal{L}_{CE} + \mathcal{L}_{dice} \tag{1}$$

ADAM optimiser [15] was utilized with a learning rate of 0.0001, β_1=0.9, and β_2=0.999 to minimize the relevant cost functions. The batch size is set to be four. Both model were trained for 60 epochs each.

Networks	Pixel Accuracy	Precision	Recall	Dice Score
Unet-vgg19	68.49	54.50	65.99	59.42
Unet-resNet34	66.85	53.25	64.31	63.85

Table 1: Performance Summary for the Networks

5 Results

The models were evaluated using overall pixel accuracy, precision, recall, and dice coefficient. All the values are obtained by averaging the results from the validation set and extra-test set. Test data were not included here since test data had no ground truth hence no statistics can be calculated. Table 1 shows the detailed quantitative performance of the networks. Both networks obtained just above 66% pixel accuracy and around 60% in terms of dice score. A slightly better recall score than precision indicates that models could handle type II error a little bit better. It is also evident that Unet-vgg19 performed better than Unet-resNet34 in terms of overall pixel-wise accuracy, the same goes for the precision and recall. However, the dice score shows the opposite picture.

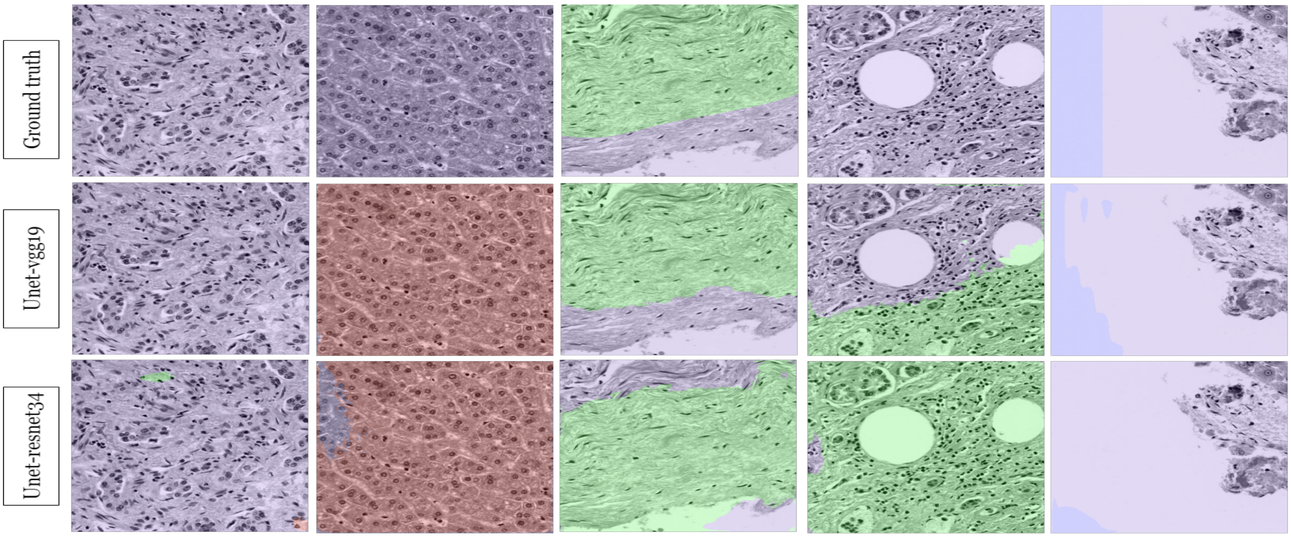

Figure 2: Sample patches from extra-test set. Ground truth (first row) and prediction mask (second row) overlayed on original images. Magenta refers to class 0 :Unknown tissue, Red refers to class 1-Parenchyma, Green refers to class 2-Fibrosis, Blue refers to Class 3-Background

Further Fig. 2 shows the qualitative performance of the models. For the first sample (first column), models were able to precisely predict the unknown tissue (class 0) type except for a few miss classified pixels by the Unet-resnet34. Still, on the following sample (second column), both models seem completely confused and predicted all the unknown type (class 0) tissue pixels to parenchyma (class 1). Unset-vgg19 was able to show decent results on the third sample (third column) by picking up nearly the same region of fibrosis (class 2) as ground-truth, but the other network had performed poorly. Finally, the predictions are partially correct on the fourth and fifth samples (fourth and fifth column), and Unet-vgg19 performed better compared to resnet based Unet.

6 Conclusion

The above quantitative and qualitative results verify the taken approach, and despite the mediocre performance, it can be concluded that the models, in general, were able to learn relevant features. The mediocre performance can be primarily explained by the lack of enough training samples and, consequently, the lack of variation in the input space. Other contributing factors in building a robust model could be adding more types of data augmentation (such as color transformation as suggested in [16], elastic deformation and Gaussian blur, etc.), experimenting with model architecture and hyper-parameter optimization.

References

[1] J. Beutel, H. L. Kundel, and R. L. Van Metter, *Handbook of medical imaging*, vol. 1. Spie Press, 2000.

[2] U. E. Lee and S. L. Friedman, "Mechanisms of hepatic fibrogenesis," *Best practice & research Clinical gastroenterology*, vol. 25, no. 2, pp. 195–206, 2011.

[3] K. Böttcher and M. Pinzani, "Pathophysiology of liver fibrosis and the methodological barriers to the development of anti-fibrogenic agents," *Advanced drug delivery reviews*, vol. 121, pp. 3–8, 2017.

[4] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[8] P. Yakubovskiy, "Segmentation models pytorch," 2020.

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[12] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, “Guided soft attention network for classification of breast cancer histopathology images,” *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1306–1315, 2019.

[13] C. Nguyen, Z. Asad, and Y. Huo, “Evaluating transformer-based semantic segmentation networks for pathological image segmentation,” *arXiv preprint arXiv:2108.11993*, 2021.

[14] Z. Jia, X. Huang, I. Eric, C. Chang, and Y. Xu, “Constrained deep weak supervision for histopathology image segmentation,” *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[15] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[16] D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, *et al.*, “Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks,” *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2126–2136, 2018.