Data Normalization, Deep Learning, and Cancer Diagnosis

Tung Dinh, Sevban Sonmez, Benjamin Price

UNIVERSITY OF HOUSTON

Department of Computer Science, University of Houston

Abstract

The ability for clinicians to accurately diagnose cancers of various types are essential for the timely and proper care of patients. However, there has been a measured increase in wait times for cancer patients since 2001. Unfortunately, a higher delay in treatment is associated with a higher chance of fatality in patients. While there are several factors associated with a delay in treatment, including gender, income, and race, one of these factors is also a delay in diagnosis. The median in wait on diagnosis time ranges from 21 days to 53.6 days based on several factors. With Convolutional Neural Networks (CNN), we can hopefully reduce the time for diagnosis. Though the obtaining and processing of images can take significant effort, we find that with one dataset the actual diagnosis of an image can take less than a second with an accuracy of over 99% using the VGG-16 model. We aim to use a variety of models using this dataset to demonstrate that the application of a CNN model to this problem should not be the barrier to improving the wait times in cancer diagnoses, and in turn, can be used to reduce cancer mortality rates, especially in highly impact communities like the poor and minority groups.

Background

Pérez-García, Sparks, and Ourselin have developed a Python library called TorchIO, providing many useful utilities for preprocessing a variety of medical images. Using techniques such as augmentation, the use of subvolumes, and the utilization of spatial metadata, data scientists can expand the size of training sets, reduce computational complexity, and correctly account for alignment and orientation of volumes in non-traditional images. In addition, Perumal and Velmurugan have shown that contrast enhancement can help doctors better view MRI images, something that can likely aid deep learning models as well,[6] and Shameena and Rahna have analyzed similar techniques for cardiac medical images. Masoudi et al. provide a quick guide on radiology image preprocessing, also acknowledging that the main barrier to alleviating this issue is not the creation of algorithms themselves, but rather it's the collection and preprocessing of image data. To demonstrate this, we attempt to use stain normalization, utilizing the technique shown by Macenko et al.

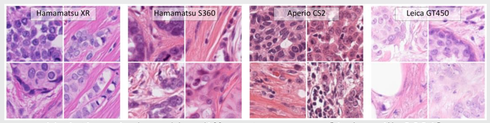
As for the CNN models we utilize, the first is a Convolutional Neural Network as covered in class. Then we use VGG-16, as pioneered by Simonyan and Zisserman. Finally we used a Residual Network, or ResNet, as developed by He, Zhang, Ren, and Sun.

Methods

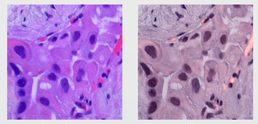
1 – PREPROCESSING

Due to dissimilarities of dye concentrations and scanner qualities during the collection of histopathological images, the data is often inconsistent, especially if they come from different labs. To overcome this, we applied the 8-step implementation of **digital stain normalization** listed on the paper "A Method for Normalizing Histology Slides for Quantitative Analysis". The process converts RGB (Red-Green-Blue) slides into Optimal Density images, and outputting optimal stain vectors.

***However, the performance deteriorated upon using the normalized images. We dropped the normalized data and only used the original data in later models.



Variations induced by different scanners [Aubreville2021]



Digital Stain Normalization

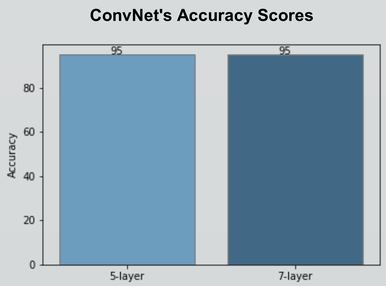
before

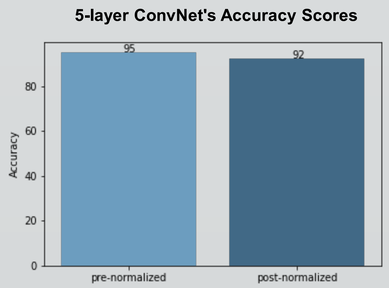
after

2 - CLASSIFICATION MODELS

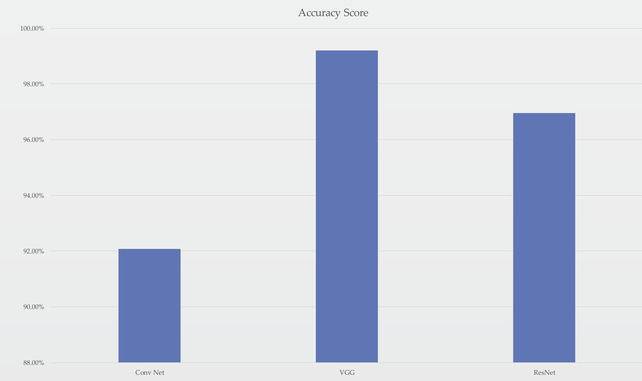
Multiple deep network architectures were used for classification:

- **5-layer ConvNet:** trained the *lung histopathological dataset* on a network with 3 convolutional layers of 3x3 filters followed by a maxpool layer of 2x2 filters, 1 dense layers of 512 nodes and the final layer of 3 nodes with a softmax classifier. Note: the lung dataset has 3 classes.
- **7-layer ConvNet:** to prevent the plunge from 512 to 3 nodes in the previous 5-layer architecture, 2 more dense layers of 126 and 60 nodes were added to the network. However, this gives little or no change to the accuracy despite an increase in the computational cost.
- **VGG-16:** trained the *colon dataset* on a deeper convolution neural network with 16 layers, including 13 convolutional layers of 3x3 filters and 3 fully-connected layers, also, a consistent use of the same padding and a maxpool layer of 2x2 filters. The consistency and uniformness of the architecture provide the most optimal accuracy of all experiments by far.
- **Resnet:** trained a *combination of both* colon and lung histopathological images on the Residual Network of 56 layers. Our goal was to find out if the dataset performs better in an even deeper architecture. However, the accuracy is slightly lower than VGG-16.





Results



We were be able to achieve an at-least 99% accuracy rating with one model, and above a 90% accuracy rating with all of our models. This will be significantly above the average accuracy of doctors, which is 71.40% on one test set used.

Conclusion

Our dataset is likely much better curated than real-world dataset examples, but the above paper found that machine learning algorithms outperformed doctors in aggregate. Doubtlessly, these algorithms can be used to supplement doctors' diagnoses and hopefully hasten them, if not outright replace them. We hope our paper contributes to the field and provides as a steppingstone for future innovations in the field of automated medical diagnostics.

Future Direction

In the future, work should focus on easing the barrier to accessing raw medical imaging data and creating a generic pipeline for feeding that data through preprocessing routines before feeding the data into training better models. Though daunting, the completion of such efforts would save lives.

Acknowledgments

We acknowledge Pérez-García, Sparks, and Ourselin, Perumal and Velmurugan, Shameena and Rahna, Masoudi et al., Macenko et al., Simonyan and Zisserman, and He, Zhang, Ren, and Sun, and Sreenivas B., and Heather Couture.