CS 231N Milestone
Predicting ISUP scores for prostate tissue samples

Josh Wolff
Stanford University
Bioengineering and Computer Science
jw1@stanford.edu

Abstract

My great-grandfather, after surviving World War II, was a victim of metastatic prostate cancer. Prostate cancer is one of the most common types of cancer in men [1]. Early diagnosis is critical for effective treatment of the disease [2]. However, misdiagnosis as prostate cancer can lead to unnecessary treatment [5]. The current standard for diagnosis is analysis of a tissue sample of the prostate. A pathologist scores the tissue viewed under the microscope on the Gleason grading system, the score of which is later converted to a number 1-5 on the ISUP scale [3]. The Gleason grading system is based on Gleason patterns, the recognition of which machine learning algorithms have demonstrated promise. Because even pathologists themselves often disagree on diagnoses for particular samples, machine learning models could provide pathologists useful recommendations in their endeavor to classify the samples, especially if such models offer an explanation for their choice [4]. The clinical assistance of such a model could lead to fewer misdiagnoses by empowering the pathologist to make a more informed decision.

1. Introduction

We consider employing computer vision techniques in deep learning to determine the ISUP score for a prostate tissue sample. In particular, we will use transfer learning and fine-tuning on well-established models. Deep learning applications in whole-slide image analysis is an emerging application of deep learning. Recently, Caie et. al. published an overview of this field, in which our problem fits nicely because the dataset consists of whole-slide images [14]. Stumpe et. al. have provided an in-depth exploration of the application of deep learning to assist in Gleason scoring, which is precisely what we aim for here [20]. Other researchers have considered deep learning in histology more broadly, which is applicable for when we fine-tune our model using other datasets in histology [18].

2. Problem Statement

We take as input a whole-slide image of a tissue sample and output an ISUP score from 0 to 5. We might also output a bounding box predicting the section of the cell most influential in its scoring, but it is not explored in the milestone.

3. Dataset

This project will use the dataset provided by the Kaggle competition, "Prostate cANcer graDe Assessment (PANDA) Challenge" [7]. This dataset consists of 11,000 whole-slide images of biopsies from two separate centers. According to the competition details, the dataset "...is the largest public whole-slide dataset available" [8]. The images are not cropped to areas of interest; however, the competition provides segmentation masks that indicate the areas of the slide influential in determining its ISUP grade.

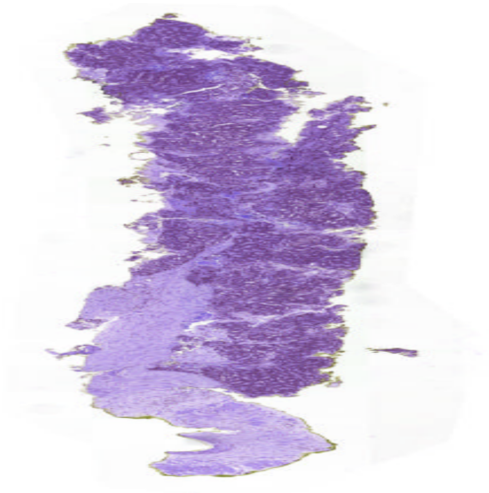

Figure 1. An example of the whole-slide image. Not shown: a segmentation mask.

For each of the models that follow, the input image was resized to 512x512 because the whole-slide images are provided in varying dimensions.

4. Technical Approach

We reduce the problem to a classification problem into classes 0, 1, 2, 3, 4, and 5.

4.1. Baseline: NN

The baseline neural network consists of flattening the input image, a fully connected layer with 1024 neurons, a sigmoid activation function, followed by an output softmax layer for classification.

4.2. Baseline: CNN

Our baseline CNN consists of 4 repetitions of a 2D convolution, batch normalization, ReLU activation, and max pooling, followed by a fully connected layer with a ReLU activation, and finally another fully connected layer with a softmax output.

4.3. Transfer Learning

Transfer learning was tested using four well-known models: SqueezeNet, DenseNet-121, ResNet-50, and Inception-v3. Each model choice was designed to give insight into how to best model the problem. Encouraged by the 0.44 quadratic weighted kappa score achieved by the baseline CNN, transfer learning using SqueezeNet was implemented with the intention of determining if a simpler, more compact model is a sufficient starting point for the task at-hand. DenseNet was tested to gain insight into the performance differential resulting from shortening the connections between the input and the final layers of the CNN [23]. ResNet-50 was tested largely because it boasts well-documented success when used as the initial pre-trained model [24]. Furthermore, we also tested Inception-v3 because Moulin et. al., who implemented deep learning for histology-focused tasks, found the Inception model to outperform VGG-16 and ResNet-50 [18].

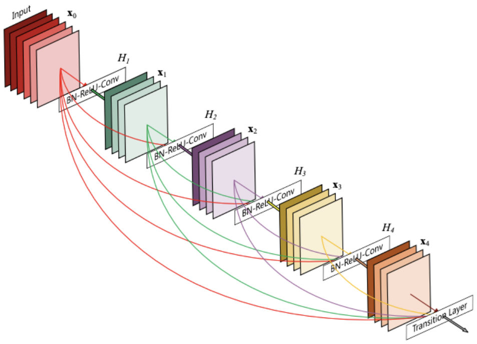

Figure 2. A visualization of the DenseNet-121 model, provided by PyTorch.com. [22]

5. Preliminary Results

5.1. Evaluation

To preliminarily evaluate the models before examining their performance on the validation set, their ability to learn the training data is first determined. To do so, the metric used is the accuracy per class. If the model is able to learn the training data, it was then evaluated on the validation test.

To evaluate the model performance on the validation set, the quadratic weighted kappa is used. This score measures the agreement between the two outcomes. This is the metric by which the entries in the competition will be judged. This measure first calculates the squared difference between the actual (i) and predicted values (j), divided by the square of the total amount of values (N) minus one.

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

The quadratic weighted kappa value is then calculated by summing the product of the weights (w_{ij}) and number of actual values that receive the predicted value (O_{ij}) divided by the product of w_{ij} and the expected outcome calculated under the assumption of zero correlation (E_{ij}).

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{i,j}}{\sum_{i,j} w_{ij} E_{i,j}}$$

This gives an intuitive result that uses the worst-case random guessing as a baseline. The stronger the correlation, the larger the κ value.

5.2. Baseline: NN

Neural networks are known to not perform well with images. For this particular class, the neural network performed exceptionally poorly, scoring 0% accuracy on classes 1 through 5, and 100% accuracy on class 0, the most common class in the training data. After multiple epochs, the network converged and was unable to improve. Furthermore, the model file consisted of 3GB of data, which is too costly to run on a remote server, which is requied for the competition. Thus, the model was not evaluated on the validation set.

5.3. Baseline: CNN

Our baseline CNN was the simplest model that was able to perform well and learn the training data. After 10 epochs, it was able to achieve 99% or greater accuracy on each of the six classes in the training set. It also provided a nice baseline for validation set performance.

Pretrained Model	Quadratic Weighted Kappa
DenseNet-121	0.61
ResNet-50	0.52
Inception v3	0.49
Custom CNN	0.44
SqueezeNet	-

Table 1. Comparison of Model Performance

5.4. Transfer Learning

SqueezeNet was unable to train on the data and unable to improve an extremely high loss. ResNet-50 and Inception v3 largely performed similarly, only slightly improving on our custom CNN. The largest gain in improvement was achieved with the DenseNet-121 pretrained model. The quadratic weighted kappa scores on the validation set are listed in Table 1.

- New results - Fine Tuning - Dataset Augmentation - Hyperparameter tuning - Class visualization methods

6. Further Work

The following sequence of steps describe the plan for further work.

1. In the original assessment of the models, each model was not subjected to the same training procedure. For example, ResNet-50 was trained for 20 epochs and DenseNet-121 was only trained for 5 epochs, and there was little justification for this difference. The models will be re-assessed using the practice of "early stopping" as the stopping point for training, where the model's training procedure is halted at the point at which the validation set performance worsens. The best-performing model will be selected for further development.

2. To alleviate the small-dataset issue, the selected model will be fine-tuned on datasets in histology, preferably in the oncology realm. The improvement as a result of fine-tuning will be measured.

3. The model's hyperparameters will then be iteratively tuned on the validation set.

4. The model will be trained on augmented data and its performance will be compared to such data.

5. Standard class visualization methods, such as a saliency map, will be employed to inform us humans what is most important in the model's classification decision.

6. If time permits, the model will use the provided segmentation masks in the competition to learn to classify multiple parts of the slide and output a bounding box for each part. Ultimately, it is conceivable that such a function would be a useful tool in a histologist's toolset in that it could enable the histologist to refocus on different areas that were perhaps overlooked.

References

References

[1] Prostate cancer. (2019, April 17). Retrieved from https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087

[2] Prostate cancer prevention and early detection. (n.d.). Retrieved from https://www.seattlecca.org/diseases/prostate-cancer/early-detection-prevention

[3] “Gleason Grading System: MedlinePlus Medical Encyclopedia.” MedlinePlus, U.S. National Library of Medicine, medlineplus.gov/ency/patientinstructions/000920.htm.

[4] Goodman, Michael, et al. “Frequency and Determinants of Disagreement and Error in Gleason Scores: a Population-Based Study of Prostate Cancer.” The Prostate, U.S. National Library of Medicine, 15 Sept. 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3339279/.

[5] McCaffery, Kirsten, et al. “Resisting Recommended Treatment for Prostate Cancer: a Qualitative Analysis of the Lived Experience of Possible Overdiagnosis.” BMJ Open, BMJ Publishing Group, 23 May 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6537980/.

[6] “Prostate CANcer GraDe Assessment (PANDA) Challenge.” Kaggle, www.kaggle.com/c/prostate-cancer-grade-assessment/data.

[7] Kaggle, https://www.kaggle.com/c/prostate-cancer-grade-assessment/overview/description

[8] Bulten, Wouter, et al. “The PANDA Challenge: Prostate CANcer GraDe Assessment Using the Gleason Grading System.” Zenodo, 19 Mar. 2020, zenodo.org/record/3715938#.XqfAjpp7lQK.

[9] ImagingLab. “ImagingLab/ICIAR2018.” GitHub, 23 Oct. 2019, github.com/ImagingLab/ICIAR2018.

[10] Litwin, Mark S, and Hung-Jui Tan. “The Diagnosis and Treatment of Prostate Cancer: A Review.” JAMA, U.S. National Library of Medicine, 27 June 2017, www.ncbi.nlm.nih.gov/pubmed/28655021.

[11] “Prostate Cancer Review.” Medscape, 23 Oct. 2018, www.medscape.com/viewarticle/903185.

[12] “Histology of Prostate Cancer.” Oncolex, Oncology Encyclopedia, oncolex.org/Prostate-cancer/Background/Histology.

[13] McNeal, J E. “Normal Histology of the Prostate.” The American Journal of Surgical Pathology, U.S. National Library of Medicine, Aug. 1988, www.ncbi.nlm.nih.gov/pubmed/2456702.

[14] Dimitriou, et al. “Deep Learning for Whole Slide Image Analysis: An Overview.” Frontiers, Frontiers, 29 Oct. 2019, www.frontiersin.org/articles/10.3389/fmed.2019.00264/full.

[15] Serag, et al. “Translational AI and Deep Learning in Diagnostic Pathology.” Frontiers, Frontiers, 30 July 2019, www.frontiersin.org/articles/10.3389/fmed.2019.00185/full.

[16] Komura, Daisuke, and Shumpei Ishikawa. “Machine Learning Methods for Histopathological Image Analysis.” Computational and Structural Biotechnology Journal, Elsevier, 9 Feb. 2018, www.sciencedirect.com/science/article/pii/S2001037017300867

[17] L., Chetan, et al. “Deep Neural Network Models for Computational Histopathology: A Survey.” ArXiv.org, 28 Dec. 2019, arxiv.org/abs/1912.12378.

[18] Sing, Tobias, et al. “A Deep Learning-Based Model of Normal Histology.” BioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2019, www.biorxiv.org/content/10.1101/838417v1.full.

[19] Nagpal, Kunal, et al. “Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer.” Nature News, Nature Publishing Group, 7 June 2019, www.nature.com/articles/s41746-019-0112-2.

[20] Eminaga, et al. “Deep Learning for Prostate Pathology.” ArXiv.org, 16 Oct. 2019, arxiv.org/abs/1910.04918.

[21] Karimi, Davood, et al. “Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images-Role of Multiscale Decision Aggregation and Data Augmentation.” IEEE Journal of Biomedical and Health Informatics, U.S. National Library of Medicine, May 2020, www.ncbi.nlm.nih.gov/pubmed/31567104.

[22] "DenseNet." PyTorch, pytorch.org/hub/pytorch_vision_densenet/.

[23] Huang, et al. “Densely Connected Convolutional Networks.” ArXiv.org, 28 Jan. 2018, arxiv.org/abs/1608.06993.

[24] Lei, Haijun, et al. “A Deeply Supervised Residual Network for HEp-2 Cell Classification via

Cross-Modal Transfer Learning." Pattern Recognition,
Pergamon, 15 Feb. 2018,
www.sciencedirect.com/science/article/pii/S0031320318300608.