Udacity Machine Learning Nanodegree

Capstone Proposal - Histopathologic Cancer Detection

Identify metastatic tissue in histopathologic scans of lymph node sections

Domain Background

Metastasis is the process by which cancer cells to spread to other parts of the body from the original cancer site. Early identification of metastasis allows an increased number of treatment options and increases the likelihood of survival [1].

A common way that cancer metastasizes is via the lymph system after moving through the walls of nearby lymph nodes. One method used to investigate the spread of cancer is histopathological analysis of sentinel axillary lymph nodes (SLNs) by a pathologist. This is a time consuming process that can be prone to error and false identification. The use of Machine Learning for image recognition can reduce manual effort, increase the chances of early recognition of cancerous cells, opening doors for more treatment options and increasing the likelihood of patient survival [2].

Problem Statement

This project will identify patches of metastic cancer in digital pathology scans from lymph node biopsies. This will aid in diagnosising is a patient's cancer. A Convolutional Neural Network (CNN) will be used to label slides as cancerous or benign.

Datasets and Inputs

The dataset is a version of the PatchCamelyon (PCam) dataset available on Kaggle [3]. The original dataset contains 327.680 color images (96 x 96px) of histopathologic scans of lymph node sections. Each image has a binary label indicating if the image contains cancerous cells [4].

The Kaggle competition training data contains 220025 labelled images, which will be used to train the model. The test data set has 57458 images and will be used to evaluate the mode. Both test and training datasets contain 50% positive and 50% negative samples.

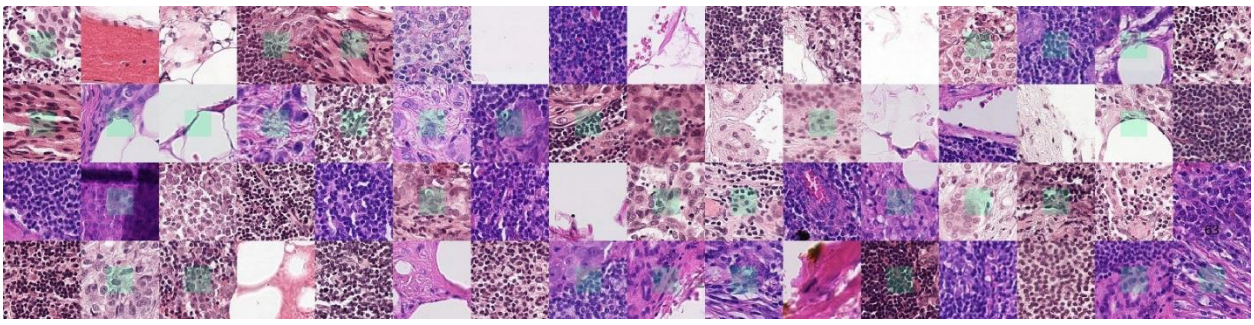

Figure 1: Example of histopathological image scans [4].

Solution Statement

The problem will be addresses by creating a CNN image classifier. Various architectures of CNNs will be explored. A fully connected layer will be used as the final output layer with a soft max activation function to predict the likelihood that the image contains cancerous cells. The model will be trained using the labelled training images.

Benchmark Model and Evalutation Metrics

In progress entries for the Kaggle competition will be used to benchmark results of the CNN. The competition uses area under the receiver operating characteristic curve (AUROC) to determine position in the leader board. The goal is to have the final model place within the top 30% of the leaderboard for the competition.

Research results looking at developing a CNN with the Camelyon16 dataset will also be used as a bench mark. The results from the study can be seen below [2]:

Name	Reference	Augmentations	Acc	AUC	NLL	FROC*
GDensenet	[1]	Following Liu et al.	89.8	96.3	0.260	75.8 (64.3, 87.2)

Project Design

The project will be completed with the following steps:

Data Exploration:

- Methods to explore, clean and filter the images will be explored.

Research and CNN Architecture exploration:

- Explore various CNN architectures
- Reading documentation on implementation of CNNs for similar use cases

Training:

- Train test split, cross validation

Testing and Optimization

- The model will be scored using the test dataset
- Model tuning

References

[1] N. C. Institute, "Metastatic Cancer," 6 Feburary 2017. [Online]. Available: https://www.cancer.gov/types/metastatic-cancer.

[2] M. V. P. J. v. D. e. a. Babak Ehteshami Bejnordi, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," JAMA, 2017.

[3] "Histopathologic Cancer Detection," https://www.kaggle.com/c/histopathologic-cancer-detection, 2018. [Online]. Available: https://www.kaggle.com/c/histopathologic-cancer-detection.

[4] "The PatchCamelyon (PCam) deep learning classification benchmark," github, [Online]. Available: https://github.com/basveeling/pcam.

[5] A. Chen, "IBM's Watson gave unsafe recommendations for treating cancer," 26 July 2018. [Online]. Available: https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science.

[6] B. Mesko, "The role of artifical intelligence in precision medicine," Taylor and Francis, Budapest, 2017.

[7] N. D. W. D. S. D. E. P. X. Zeya Wang, "Medium," 6 August 2018. [Online]. Available: https://medium.com/@Petuum/deep-learning-for-breast-cancer-identification-from-histopathological-images-f38de0a658a5.

[8] T. P. K. P. M. V. D. I. Konstantina Kourou, "Machine learning applications in cancer prognosis and prediction," Science Direct, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037014000464.