Udacity Machine Learning Engineer Capstone Project Proposal: Histhopathologic Cancer Detection

Arief Ramadhan

January 2020

1 Domain Background

The field of digital pathology has facilitated image analysis algorithms to assist and automate diagnostic tasks [1]. A proven approach is to use Convolutional Neural Networks (CNNs), a type of deep learning model which is most commonly applied to analyze image. CNN has been shown to outperform pathologists in a variety of diagnostic tasks [1].

This project aims to identify metastatic cancer in small image patches taken from larger digital pathology scans [1]. Early detection and treatment of cancer play a crucial role in patient's survival rate. By the time symptoms appear, cancer may have begun to spread and be harder to treat. We use CNN to predict whether an image contains at least one pixel of tumor tissue [1], therefor help improve early diagnosis of cancer.

2 Problem Statement

Given images of digital pathology scans, determine whether there is at least one pixel of tumor tissue in the images. We approach the problem as a binary classification of a "tumor(1)" class and a "no tumor(0)" class.

3 Datasets and Inputs

The dataset is provided by Kaggle for a past competition [2]. It consists of a large number of small pathology images to classify. The link to the dataset can be found here.

4 Solution Statement

We will solve this problem by using Convolutional Neural Network that has proved to be successful in classifying image. The solution is a binary classification model capable of predicting whether an image contains tumor or not. We might use transfer learning to speed up the model's training.

5 Benchmark Model

A work in [1] found that rotation equivariance significantly improves tumor detection performance on a challenging lymph node metastases dataset. A similar method might be adapted in our solution.

We find a variety of projects on Github that solved this particular problem with different strategies, as displayed in Table 1.

Table 1: Past Projects on Github

Author	Model	Accuracy
G. Montamat	Transfer Learning (Resnet50) + MLP	cell6
G. Surma	Transfer Learning + CNN	95.8%
R. Wagulde	Ensemble Learning + CNN	cell6

6 Evaluation Metrics

Our model is optimized using only one metric, accuracy, and evaluated with it and two other metrics, precision and recall. The short explanations of each of metric, for our case, are:

- Accuracy: the number of correct classification divided by the total number of classification
- Precision: the number of correct tumor detection divided by the total number of tumor detection
- Recall: the number of correct tumor detection divided by the number of tumor images

7 Project Design

Our model is built using Python with Pytorch, Numpy, Panda, and Sklearn libraries.

7.1 Workflow

The project's workflow:

- Explore and understand dataset
- Preprocess dataset. The obvious one is to crop the images into 32x32 pixels size.
- Build the CNN Model
- Hyperparameter tuning
- Evaluate the result

References

[1] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *International Conference on Medical image computing and computer-assisted intervention.* Springer, 2018, pp. 210–218.

[2] Kaggle. (2019) Histopathologic cancer detection. [Online]. Available: https://www.kaggle.com/c/histopathologic-cancer-detection/overview