Machine Learning Engineer Nanodegree

Capstone Proposal

Chenxin Wang
February 23rd , 2019

Histopathologic Cancer Detection

Domain Background

Histopathologic cancer detection has been performed by human pathologists reviewing stained specimen on slide glasses with microscopes. This process is time-consuming and error-prone. A recent technique, which digitizes glass slides into whole-slide images (WSIs), facilitates image analysis with machine learning algorithms to assist diagnostic tasks. Computer assisted diagnosis, if successful, can significantly reduce the workload of pathologists and allow them to focus on more difficult diagnosis.

Among these machine learning algorithms, deep convolutional neural networks (CNNs), trained on whole-slide images, have shown great prospect in the cancer detection. In Camelyon16 [1], a grand challenge to identify top-performing machine learning algorithms for the automated detection of metastatic breast cancer, the winner team [2] trained two deep neural networks (GoogLeNet), one on the whole training sets and the other on harder examples, and took an average of the two models' predictions as the cancer probability. They achieved an area under the receiver operating curve (AUC) of 0.925 for the whole slide image classification and a score of 0.7051 for the cancer localization task. This team later improved these metrics to 0.994 and0.807 respectively using color normalization [3] and additional data augmentation. Liu et.al. [4] reached a score of 0.924 for the cancer localization task by fine-tuning an Inception-V3 deep neural network model on the Camelyon16 dataset.

Problem Statement

This project aims to create a deep learning algorithm to identify metastatic cancer in small image patches taken from larger whole-slide images. The algorithm is trained on a set of labeled images, and evaluated on a test set. Each training image patch is labeled 1 if it has tumor and 0 if not. Given an image patch, the algorithm predicts tumor probability.

Datasets and Inputs

The dataset comes from a Kaggle competition [5]. It has 220,026 pathology images in the training set and 57,459 pathology images in the test set. Each image in the training set is labeled with 1 or 0. The label 1 indicates that the center 32x32px region of an image contains at least one pixel of tumor tissue. We will predict probabilities for images in the test set and submit to Kaggle for evaluation.

Solution Statement

The training set will be split into one larger set for training and one smaller set for validation. Some data augmentation techniques like image rotation and color perturbation will be implemented to combat the rarity of tumor patches. I will first try a CNN from scratch to see how accurate it can reach. Next, I will try transfer learning, fine-tuning some well-known architectures such as ResNet50, DenseNet169 and NasNet. At the end, I will focus on one architecture that gives the highest AUC and tune parameters to improve it.

Benchmark Model

The top 50% in Kaggle public leaderboard achieves an AUC of 0.961, which may change because the competition is still alive. I will use this score as a benchmark.

Evaluation Metrics

In this Kaggle competition, solution models are evaluated on area under the ROC curve between the predicted probability and the observed target on the test set. Since the ground truth for the test set is not given, we need to submit predictions to Kaggle for evaluation.

Before the definition of the area under the ROC curve is given, several terms are defined in the following.

- TP: True Positive
- FN: False Negative
- FP: False Positive
- TN: True Negative
- TPR (True Positive Rate) = TP/(TP+FN)
- FPR (False Positive Rate) = FP/(TN+FP)

The TPR and FPR are evaluated as the threshold probability dividing the positive and negative varies. The ROC curve is plotted with the TPR against the FPR with the TPR as the y-axis and the

FPR as the x-axis. The area under the ROC curve represents the performance of a solution model. An excellent model has the area under the ROC curve close to 1. When the area under the ROC curve is 0.5, the model has no classification capability. When the area under the ROC curve is close to 0, the model predicts the opposite of the ground truth.

Project Design

The following workflow is proposed:

- Data Pre-processing
    - Understand the distribution of training labels
    - Data augmentation
- Training
    - Train a CNN from scratch
    - Transfer learning with ResNet50, DenseNet169 and NasNet
- Fine-tuning
    - Tune parameters for a selected architecture to improve the score

Reference

[1] Camelyon 2016. https://camelyon16.grand-challenge.org/.

[2] Wang, D., et al.: Deep learning for identifying metastatic breast cancer. (2016)

[3] Bejnordi, B.E., et al.: Stain specific standardization of whole-slide histopathological images. IEEE Trans. on Medical Imaging 35(2), 404–415 (2016)

[4] Liu, Y., et al.: Detecting cancer metastases on gigapixel pathology images. (2017)

[5] Kaggle competition. https://www.kaggle.com/c/histopathologic-cancer-detection.