Machine Learning Engineer Nanodegree

Capstone Proposal

Adrian Voicu
October 06, 2020

Digital Pathology - Mitosis Detection using Transfer Learning in PyTorch

Domain Background

Digital pathology (DP) is a subdomain of pathology that focuses on the analysis and processing of digital scans of microscope samples. The progress of scanning and microscopy technologies enables expert pathologists to visualize and share high resolution images using computers or even mobile devices.

According to Leica Biosystems, a company producing DP tools, the advances in this field will ultimately allow computer aided diagnostics and personalized medicine to be made available to patients. [4]

Standing in the broader field of DP, is the specific task of Mitosis Detection (MD).
In the context of breast cancer tumor diagnosing, MD refers to the number of self dividing cells. A high number of mitotic centers is correlated with an advanced stage of the tumor. The correct assessment of this indicator is extremely important in selecting a patient's treatment plan (ie. a patient with a less pronounced tumor might be spared the aggressive treatment that a patient with a more developed tumor might require).[2]

Mitosis Detection is a task currently done manually by trained specialists in a pathology lab: glass slides of biopsy samples extracted from the patient are analyzed by the pathologist and assigned a grade, based on specific characteristics observed in the sample (in short, the number of mitotic centers detected). The grade is then used in assigning a personalized treatment to the patient. Thus, a correct and thorough interpretation of the biopsy scans is vital. The challenges involved in this workflow are not few: the subjective nature of the results (the cells can have different shapes, depending on their development stage), the high volume of work and attention needed for analyzing even a single slide, the decreasing number of trained specialists, the increasing number of patients requiring a diagnostic. [4]

It can be stated that MD constitutes a field where the development of an automated, reproducible and reliable process can be highly beneficial. However, the desired outcome of automation is not to completely replace an expert analysis, but merely to provide a second trustworthy opinion.

This work is inspired by the TUPAC16 challenge [2]. A strong personal motivation also came in play, due to the believe that using ML technologies can be beneficial and directly improve human life. One source of inspiration was also the impressive story of the Biotech Startup OncoStem, that strive not only to improve the existing methods, but to make them available to people and countries with less financial power. [3]

Problem statement

The problem addressed in this project is a **binary image classification**. The model must be trained using real life laboratory images and must predict with a reasonable degree of confidence whether an image contains a mitotic center or not. The images annotated by pathology experts will be used as the ground truth.

As shown by [1] and others, Deep Learning techniques have been proved to give good results in Digital Pathology tasks such as Mitosis counting and detection.
The goal of this project is therefore to implement a basic solution, that can be used as a starting point for further research and improvement.

Datasets and inputs

The **dataset** used for the project shall be the publicly available set of images found on Andrew Janowczyk's blog [1], which includes a list of 311 images (.tif format), with 550 mitotic centers expertly annotated (coordinates of the centers included in the .csv file). These are high resolution scans of microscopic images containing biopsy samples from different patients.
Attached to each .tif image (which represents an original H&E[1] lab scan), there will be 2 extra helper images:

- one containing the dilated Blue Ratio Segmentation[2] (BRS) mask of the original (*_pc.png – further referenced as "type 1" input image) and
- one containing circles drawn around the mitotic circles overlaid on the BRS mask (*_cmask.png – further referenced as "type 2" input image).

Solution Statement

The proposed solution tries to follows the approach from source [4], but is implemented using the **Pytorch** framework and an **all-Python** implementation.

Since we are dealing with high resolution images that need to be broken down in small parts, a major part of the implementation will be constituted by the pre-processing of the input data. After pre-processing, the dataset is expected to contain tens of thousands of images, which represent both positive and negative samples
A transfer learning method will be used, using AlexNet as a pre-trained model.

Benchmark model

The expected result is an accuracy of more than 50%. Results from [2] and [4] can be used for evaluating the final outcome. Having in mind that this is a highly complex issue, which requires a skill set which is outside the scope of this nanodegree, the accuracy is not expected to be at cutting edge levels. The focus is mainly directed towards developing a working pipeline that can be further improved and built upon, but also getting familiar to the challenges of a real-life ML problem.

---

1 Hematoxylin and Eosin

2 Using the BRS technique is useful to isolate the patches where mitotic centers are most likely to be found. The BR centers are dilated, in order to include the centers context in the processing. For more details on BRS technique, see [6]

Evaluation metrics

Since the problem represents a binary classification, the main metric used for evaluating the model will be the **accuracy** measurement. A subset of the original dataset will be extracted and used solely for testing and evaluating the model. Accuracy measurement will be made on extracted patches (a scan can contain multiple positive patches) from the original image. The class of the patches is set according to their presence in the type 2 images (ie. class 1 – patch is present in the type 2 image, representing annotated mitotic centers; class 0 – class is not present in the list of annotated mitotic centers)

Project Design

One of the big challenges of the project is the patch extraction phase that needs to be done on the original dataset. This is a highly resource consuming operation.
The workflow will follow the procedure briefly described below:
- for extracting positive patches: from the original image (*.tif) extract the annotated mitotic centers by using the centers found in the type 2 input images.
- for extracting negative patches: from the original image (*.tif) extract the annotated mitotic centers by using the centers found in the type 1 input images. This makes sure that the negative dataset contains patches that are relevant for the classification and not just random patches from the scan that contain no annotation.

To address the problem of class imbalance, the number of positive patches will be augmented by applying a rotation to each extracted patch.

The AlexNet model was chosen, since it's a benchmark network that has a straightforward implementation in PyTorch. The transfer learning approach was selected, since it is shown to give good results with a smaller resource expenditure compared to the "clean"(untrained) model.

In order to use the pre-trained AlexNet model, a set of transforms needs to be applied to the images obtained after the initial patch extraction phase(eg. reshaping, converting to tensors, normalization, creation of DataLoaders). These are relatively easy to implement using the PyTorch framework, which has built-in functions for most of the needed operations.

Since the training of the model can still be quite time consuming (when using no GPU), it is also important to implement a solution that can handle a fragmented workflow (ie. stop training, saving and loading the model and it's trained weights, resume training).

Optionally, a second pre-trained model can be trained (eg. ResNet, DenseNet), for comparing the performance of the two.

References

[1] Use Case 5: Mitosis Detection. http://www.andrewjanowczyk.com/use-case-5-mitosis-detection/

[2] Tumor Proliferation Assessment Challenge 2016. http://tupac.tue-image.nl/

[3] How This Biotech Company Integrates Machine Learning Algorithms To Detect Relapse Of Breast Cancer. https://analyticsindiamag.com/how-this-biotech-company-integrates-machine-learning-algorithms-to-detect-breast-cancer/

[4] Digital Pathology. https://www.leicabiosystems.com/knowledge-pathway/digital-pathology/

[5] Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images. Wahab, N., Khan, A., & Lee, Y. S. (2019).

[6] Wang H, Cruz-Roa A, Basavanhally A, Gilmore H, Shih N, Feldman M, et al. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. J Med Imaging (Bellingham) 2014;1:034003.