Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Program

Kaggle Competition Histopathologic Cancer Detection

Machine Learning final project report

Authors:
Oleh Lukianykhin
Yevhen Pozdniakov
(contributed equally, order chosen alphabetically)

March, 2019

1 Introduction

Computer Vision (CV) is an important part of the modern Data Science. It applies interdisciplinary knowledge to the problem of gaining high-understanding by computer from digital images or videos.

It has achieved a lot of impressive results in different application fields, such as automotive, security, manufacturing, quality control and healthcare. For instance, CV research results were applied to develop text scanners, automatic inspections at factories, detect faces and classify different objects on photos. Also they are utilized in self-driving cars and robot control. CV results are so good, that it is used to detect signs of certain illnesses.

In general, application of CV allows to automate certain routines and avoid mistakes caused by a human factor.

Thus we would like to focus on application of Machine Learning methods to a CV problem in our final project.

2 Motivation

Excluding human factor is especially important, when one tries to decrease inefficiency in healthcare. For example, wrong prescriptions because of doctors fatigue or inexact diagnoses lead to a longer recovery time, medicines overuse and, in some cases, patients death.

Thus by applying CV methods money and people's lives can be saved. In particular, we want to focus on cancer detection problem.

Cancer kills a lot of people every year, e.g. in 2018 9.6 million people died worldwide because of different types of cancer. It is in the list of top mortality reasons by World Health Organization. Early detection helps to decrease mortality rate. So by developing CV method for detecting cancer at the early stage from images of tissues samples, significant impact can be achieved.

Moreover, solution of this CV problem has a potential to help in solving other problems, as this solution can be applied to other tasks.

3 Problem Setting

As this project was inspired by Kaggle competition, data, exact formulation and other requirements, such as quality metric, were taken from it.

3.1 Problem Formulation

The task is to develop a model that detects cancer cells at histopathological images. These are microscope images of tissues' samples from patients lymph nodes. Samples are stained with special chemical markers.

According to the competition requirements, cancer metastases should be detected only in the central region of the image. Other parts of the image can be used as additional information for detection.

Hence, it is a classic binary classification problem.

3.2 Data

Dataset of labeled images was provided by competition organizers. Images have size 96x96 pixels. Size of the central region, in which cancer cells should be detected, is 32x32 pixels.

Dataset contains 220 025 unique labeled images. Positive and negative classes are claimed to be balanced in the data.

This dataset is modified version of the PatchCamelyon (PCam) [10],[3] benchmark dataset. In particular, for the competition duplicates were removed,

3.3 Metric

According to the competition requirements ROC-AUC metric should be used. Because of that we will continue using it in our work.

4 Related Work

Histopathologic cancer detection can be considered as a typical image recognition problem. A proven approach is to use convolutional neural networks (CNNs) and deep learning which have successfully increased the performance of such analyses [4]. This approach has been shown to outperform pathologists in a variety of tasks [3]. However deep learning models require significant amount of an annotated data, so transfer learning conception is commonly used [7].

There have been a lot of studies based on transfer learning with CNNs, to detect cancerous regions [7], [2]. In general, the following approaches are used:

- use a model that has performed well in other image recognition tasks and to train it from scratch [8];
- use a pre-trained model as a feature extractor and perform further classification with a separate classifier [5], [9].
- use a pre-trained model and fine-tune it with the data at hand, in other words freezing the initial neural network layers and fine-tuning the last few layers [1], [2].

Huge variety of models (with small modifications) are suitable for histopathologic images detection: AlexNet, VGG, ResNet, InceptionV3, MobileNet and DenseNet [6]. Most of the transfer learning schemes use architectures or models trained on ImageNet dataset [6].

A model based on the observation that histopathology images are inherently symmetric under rotation and reflection was proposed in [10]. According to this paper rotation equivariance improves reliability of the model and general tumor detection performance.

5 Approach to solution

One of the purposes of this project is to get familiar with the CV techniques. So now we will describe just a general approach to the solution.

According to the related work overview Convolutional Neural Networks are capable to solve problems, like this. Thus, we would like to start with their application to this problem. Applicable architectures were described in the previous section.

Although, data can be used as is, data preparation techniques can be applied. In particular, we may to start from cropping image to the size of region, where cancer cells should be detected (32x32 pixels). Then, other crop sizes can be considered - various amount of additional information for model training will be available.

Data augmentation combined with test time augmentation may improve prediction quality as well.

References

[1] Eirini Arvaniti and Manfred Claassen. Coupling weak and strong supervision for classification of prostate cancer histopathology images. *arXiv preprint arXiv:1811.07013*, 2018.

[2] Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, and Manfred Claassen. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8, 2018.

[3] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[4] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. Deep learning for image-based cancer detection and diagnosis- a survey. *Pattern Recognition*, 83:134–149, 2018.

[5] Hanna Källén, Jesper Molin, Anders Heyden, Claes Lundström, and Kalle Åström. Towards grading gleason score using generically trained deep convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1163–1167. IEEE, 2016.

[6] Umair Akhtar Hasan Khan, Carolin Strenberg, Oguzhan Gencoglu, Kevin Sandeman, Timo Heikkinen, Antti Rannikko, and Tuomas Mirtti. Improving prostate cancer detection with breast histopathology images, 2019.

[7] Brady Kieffer, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.

[8] Kunal Nagpal, Davis Foote, Yun Liu, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, Greg S Corrado, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *arXiv preprint arXiv:1811.06497*, 2018.

[9] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[10] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.