Bachelor’s thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Automatic Detection of Metastases in Whole-slide Lymph Node Images Using Deep Neural Networks

Pavlína Koutecká

Supervisor: prof. Dr. Ing. Jan Kybic
Field of study: Cybernetics and Robotics
August 2020

BACHELORÓS THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Koutecк§ Pavl²na** Personal ID number: **474383**
Faculty / Institute: **Faculty of Electrical Engineering**
Department / Institute: **Department of Cybernetics**
Study program: **Cybernetics and Robotics**

II. BachelorÓs thesis details

BachelorÓs thesis title in English:

Automatic Detection of Metastases in Whole-Slide Lymph Node Images Using Deep Neural Networks

BachelorÓs thesis title in Czech:

Automatick§ detekce metast§z v histologickĨch obr§zc²ch lymfatickĨch uzlin pomoc² hlubokĨch neuronovĨch s²t²

Guidelines:

Develop a method based on deep convolutional neural networks for solving the task of the detection of metastases in whole-slide lymph node images using deep neural networks, as defined in the Kaggle Histopathological Cancer Detection, CAMELYON16 and CAMELYON17 challenges. Get familiar with related work from the literature and develop a baseline solution for patch classification using the ResNet architecture and test it on the data from the Kaggle Histopathological Cancer Detection challenge. Improve and extend these techniques for the full slide segmentation as required by the CAMELYON16 challenge. Implement a slide-level aggregation. Evaluate the performance using the CAMELYON16 criteria. Consider possible improvements using for example attention networks or machine-learning based aggregation.
Time-permitting, consider implementing the patient-level aggregation as defined by the CAMELYON17 challenge and end-to-end learning based on the patient-level weakly labeled data.
Evaluate your results experimentally on the provided datasets and submit your solution to the above mentioned online challenges to compare the performance of your method with state of the art.

Bibliography / sources:

[1] Kaggle. Histopathologic Cancer Detection. https://www.kaggle.com/c/histopathologic-cancer-detection.
[2] The CAMELYON17 challenge. https://camelyon17.grand-challenge.org/.
[3] The CAMELYON16 challenge. https://camelyon16.grand-challenge.org/.
[4] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318(22):2199ï2210. doi:10.1001/jama.2017.14585
[5] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H. Beck. Deep Learning for Identifying Metastatic Breast Cancer. http://arxiv.org/abs/1606.05718

Name and workplace of bachelorÓs thesis supervisor:

prof. Dr. Ing. Jan Kybic, Biomedical imaging algorithms, FEE

Name and workplace of second bachelorÓs thesis supervisor or consultant:

Date of bachelorÓs thesis assignment: **09.01.2020** Deadline for bachelor thesis submission: **14.08.2020**

Assignment valid until: **30.09.2021**

prof. Dr. Ing. Jan Kybic SupervisorÓs signature	doc. Ing. Tom§ġ Svoboda, Ph.D. Head of departmentÓs signature	prof. Mgr. Petr P§ta, Ph.D. DeanÓs signature

Acknowledgements

Foremost, I would like to express my sincere appreciation to my supervisor, prof. Dr. Ing. Jan Kybic, for a tremendous amount of patience and great guidance throughout the whole process of making this thesis. He always motivated me to give my best and was great support whenever I ran into trouble.

I wish to show my gratitude to the people whose assistance also helped me with the completion of this thesis. Dr. ret. nat. Jan Hering for his technical advice which helped me to manage the thesis, and doc. MUDr. Tomá² Ku£era, Ph.D. for his medical guidance and valuable insights.

Besides these people, I would like to thank Center for Machine Perception for the opportunity to be part of it and use its technical and hardware resources.

Finally, I wish to express my deepest gratitude to my family and friends for providing me with unfailing love and continuous encouragement throughout my whole life and especially the study years. Without you, this accomplishment would not have been possible. Thank you.

Declaration

I hereby declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the Methodical instructions for observing the ethical principles in the preparation of university thesis.

Prague, 14. August 2020

Abstract

Digitisation of cancer recognition in histopathological images is researched topic in recent years, and automated computerised analysis based on deep neural networks has shown potential advantages as a diagnostic strategy. In this thesis, we develop a method for solving the task of automatic metastases detection in whole-slide lymph node images. We are motivated mainly by three existing grand challenges from the histopathologic area: Histopathologic cancer detection challenge by Kaggle, CAMELYON16 and CAMELYON17. First, the baseline solution using ResNet-50 architecture is developed in order of solving the patch classification as defined in Kaggle's challenge. Baseline solution is then extended, and the method is improved to perform the task of tumour segmentation. We propose to use DeepLabV3 architecture and compare it with Fully Convolutional Network and UNet architectures. DeepLabV3 proves to be the most capable model for tumour segmentation. Slide-level and patient-level aggregation are implemented using two classifiers Random forest and XGBoost. The evaluation shows that their performance is comparable.

The proposed solution is tested and uploaded to the above mentioned grand challenges. For all three challenges, our solution proves to be competitive among other participants.

Keywords: deep learning, machine learning, pathology, breast cancer, classification, segmentation, biomedical imaging, neural network

Supervisor: prof. Dr. Ing. Jan Kybic

Abstrakt

Digitalizace procesu detekce rakoviny v histopatologických snímcích je p°edm¥tem výzkumu posledních let a automatizovaná po£íta£ová analýza zaloºená na hlubokých neuronových sítích ukázala potenciální výhody jako diagnostická strategie. V této práci vyvíjíme metodu pro °e²ení úlohy automatické detekce metastáz v histologických snímcích lymfatických uzlin. Motivací jsou zejména tyto t°i existující sout¥ºe z histologické oblasti: sout¥º v detekci rakoviny od Kaggle, CAMELYON16 a CAMELYON17. Nejd°íve je vyvinuto základní °e²ení vyuºívající architekturu ResNet-50 pro klasifikaci patch·, stejn¥ jako je definováno v Kaggle sout¥ºi. Toto °e²ení je poté roz²í°eno a metoda je vylep²ena tak, aby provád¥la segmentaci nádor·. Navrhujeme pouºití architektury DeepLabV3 a její porovnání s architekturami Fully Convolutional Network a UNet. DeepLabV3 se ukazuje jako nejschopn¥j²í model pro segmentaci nádor·. Následná agregace na úrovni snímk· a na úrovni pacient· je implementována pomocí dvou klasifikátor· - Random forest a XGBoost. Evaluace ukazuje, ºe výkon obou klasifikátor· je srovnatelný.

Navrºené °e²ení je otestováno a nahráno do vý²e uvedených sout¥ºí. Pro v²echny t°i sout¥ºe se na²e °e²ení ukázalo jako konkurenceschopné.

Klí£ová slova: hluboké u£ení, strojové u£ení, patologie, rakovina prsu, klasifikace, segmentace, biomedicínské zobrazování, neuronová sí´

P°eklad názvu: Automatická detekce metastáz v histologických obrázcích lymfatických uzlin pomocí hlubokých neuronových sítí

Contents

List of abbreviations and acronyms

AI	Artificial Intelligence
API	Application Programming Interface
ASAP	Automated Slide Analysis Platform
AUC	Area Under the ROC Curve
BACH	BreAst Cancer Histology images
BW	Black and White
CAMELYON16	CAncer MEtastases in LYmph nOdes challeNge 2016
CAMELYON17	CAncer MEtastases in LYmph nOdes challeNge 2017
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
CWZ	Canisius-Wilhelmina Hospital
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
FCN	Fully Convolutional Network
FN	False Negative
FP	False Positive
FROC	Free-response Receiver Operating Characteristic
H&E	Hematoxylin and Eosin
HMS	Harvard Medical School

HSI	Hue-Saturation-Intensity
HSV	Hue-Saturation-Value
IoU	Intersection over Union
ITC	Isolated Tumor Cells
LPON	Laboratorium Pathologie Oost-Nederland
MGH	Massachusetts General Hospital
MIT	Massachusetts Institute of Technology
ML	Machine Learning
PANDA	Prostate cANcer graDe Assessment
PCam	PatchCamelyon
pN-stage	pathologic N stage
QWK	Quadratic Weighted Kappa
RGB	Red-Green-Blue
ROC	Receiver Operating Characteristic
RST	Rijnstate HoSpiTal
RUMC	Radboud University Medical Centre
TIFF	Tagged Image File Format
TN	True Negative
TNM	Tumor, Nodes, Metastasis
TP	True Positive
TUPAC16	TUmor Proliferation Assessment Challenge 2016
UMCU	University Medical Centre Utrecht
WHO	World Health Organization
WOTC	WithOut a Time Constraint
WSI	Whole Slide Image
WTC	With a Time Constraint
XML	eXtensible Markup Language

Chapter 1

Introduction

According to the World Health Organization (WHO), breast cancer is the most frequently diagnosed cancer in Czech women. In 2018, there were 7 436 newly diagnosed patients with breast cancer, which accounts for 25 % of all diagnosed women cases, and 1 580 patients died from this disease [1]. In the present research, histopathologic image analysis is the standard method applied in the clinical practice to diagnose breast cancer. Even though the prognosis for patients diagnosed with breast cancer is usually good, the survival rate declines if cancer metastasises [2]. That makes recognising the metastases in lymph node sections one of the most important prognostic factors.

In the process of histology image analysis for cancer diagnosis, pathologist standardly visually observes the tissue, its distribution and regularities of cell shapes. After that process, pathologist decides whether there are some cancerous tissue regions and determines the malignancy level [3]. However, this diagnostic procedure is time-consuming and small metastases are very difficult to detect even for experienced pathologists [4]. Fortunately, computer-based image analysis has become a rapidly expanding field within the past few years [3] and whole-slide scanners are now commonly used for digitising glass slides at high resolution. This process partially allows automation of the histopathologic image analysis for cancer diagnosis, but there is still a great potential to improve and fully automate this task and help the pathologists to reduce their workload.

1.1 Motivation

Considering the recent improvements in the field of machine learning (ML) algorithms and whole-slide imaging, the task of fully automated analysis of histopathologic images started to be more approachable than ever. The availability of many digitalised whole-slide images resulted in increasing interest of the medical image analysis community, and numerous histopathologic imaging challenges in cancer diagnosis arose lately to improve the efficiency and accuracy of this task. Commonly, a clinically relevant task, like cancer detecting or grading, predicting prognosis or identifying metastasis, is defined by organisers, who provide a sufficiently comprehensive and diverse collection of data called dataset Participants use the dataset to develop an ML algorithm appropriate for the specified task, which is subsequently evaluated by the challenge organisers. Typically, the submission deadline follows a workshop or conference, where participants with best-scored algorithms discuss their approaches and solutions. This procedure led to quick progress in automated histopathology image analysis and allowed a meaningful comparison of algorithms with promising results.

Many successful medical imaging challenges were organised in recent years. In histopathology field, it was, for example, breast cancer histology images challenge (BACH) [5], tumour proliferation assessment challenge (TUPAC16) [6] and ongoing prostate cancer grade assessment challenge (PANDA) [7]. This thesis is mainly motivated by three existing challenges Histopathologic cancer detection challenge by Kaggle [8 10], Cancer metastases in lymph nodes challenge 2016 (CAMELYON16) [10] and Cancer metastases in lymph nodes challenge 2017 (CAMELYON17) [11].

1.1.1 Histopathological cancer detection challenge by Kaggle

This challenge[1] aims to create an algorithm to identify metastatic cancer in small image patches taken from large digital pathology scans. The data for this challenge is a slightly modified version of the PatchCamelyon (PCam) dataset, which was derived from the CAMELYON16 dataset [8 10]. Kaggle runs this competition since 2019.

The task of this competition is very straight-forward a clinically-relevant task of the metastasis detection is presented as a binary image classification task. Models for this task are easily trainable in a couple of hours, and its performance is evaluated on the area under the receiver operating characteristic (ROC) curve. That makes this competition an excellent resource for fundamental research on topics as digital pathology, automatic tumour detection and whole-slide imaging.

---

[1] Available at https://www.kaggle.com/c/histopathologic-cancer-detection/ .

1.1.2 CAMELYON16 challenge

The goal of this challenge[2] is to develop an algorithm for automated detection of metastases in whole-slide images of lymph node sections [10]. Two medical centres in the Netherlands provided an extensive dataset. This competition consists of two tasks [10]:

1. Slide-based evaluation: Algorithms are evaluated for their ability to discriminate every whole-slide image as either containing or lacking metastases. For the evaluation, the ROC curve is used.
2. Lesion-based evaluation: Algorithms are evaluated for their ability to identify individual micro-metastases and macro-metastases in whole-slide images. For the evaluation, the free-response receiver operating characteristic (FROC) curve is used.

Different evaluation metrics for every task resulted in two independent algorithm rankings. Challenge was opened to new entries only in 2016.

1.1.3 CAMELYON17 challenge

The goal of this challenge[3] is, same as in CAMELYON16 challenge, to develop an algorithm for automated detection of metastases in whole-slide images of lymph node sections. Compared to the CAMELYON16 challenge, the dataset is notably extended data were provided by five medical centres. Challenge is open to new entries since 2017.

The task of the competition developed from slide-level analysis to patient-level analysis. In this challenge, artificial patients are created. There are five slides provided for each patient, and every slide corresponds to one lymph node section. This approach combines the detection and classification of metastases in multiple lymph node slides, assigned to one patient, into one outcome corresponding to the patient pN-stage, closely described in Chapter 2 [11]. This brings the task closer to clinical practice. Usually, many lymph node slides are prepared for the patient, and aggregating results of more slides is a necessary step to involve an algorithm for automated detection of metastases in daily medical practice. For the evaluation of the results, the five-class quadratic weighted kappa is used.

---

[2] Available at https://camelyon16.grand-challenge.org/.

[3] Available at https://camelyon17.grand-challenge.org/.

1.2 Goals

The main focus of this work is to develop a method for solving the task of the detection of metastases in whole-slide lymph node images using deep convolutional neural networks (DCNNs), as defined in the Kaggle Histopathological Cancer Detection, CAMELYON16 and CAMELYON17 challenges. To achieve that, it is necessary to get familiar with related work from the literature and current state-of-the-art methods.

In the following chapters, a baseline solution for patch classification using deep neural networks (DNNs) will be created and tested on the data from the Kaggle Histopathological cancer detection challenge. This technique will be improved, and patches will be aggregated to provide the full slide segmentation and slide-level classification as required by the CAMELYON16 challenge. The patient-level aggregation will extend the slide-level solution as required by the CAMELYON17. Both slide-level and patient-level results will be evaluated experimentally on provided datasets, and the final solution will be submitted to the CAMELYON17 challenge to compare the performance of our method with state-of-the-art.

Moreover, some parts of this work will be expanded with additional information from the medical field to analyse the problematics comprehensively, localise weaknesses of our method and provide the reader with a better understanding of the medical background.

Chapter 2

Medical background

2.1 Anatomy of the breast

As different parts of the breast will be referenced repeatedly, a better understanding of its anatomy will help us deal with the task. A healthy female breast, shown in Figure 2.1, consists of 15 to 20 globes of glandular tissue, called *lobes* [13]. Each of the lobes is made up of smaller *lobules* glands that produce milk. These lobules are arranged in clusters, similarly as grapes, and connected by milk ducts, which carry the milk to the nipple [14]. Lobes are supported by the *fibrous connective stroma* forming a latticed framework, travelling through the breast and inserting into the dermis. That provides remarkable mobility while still supporting the breast [13, 15]. The remainder of the breast is formed by fat cells called *adipose tissue*, which fills the space between the lobes and fibrous stroma. Breast cancer typically starts to form

(a) : Front view

(b) : Side view

Figure 2.1: Detailed illustration of the adult female breast anatomy, taken and edited from [12].

in the structure of lobes and ducts [16].

2.2 Lymphatic system

The lymphatic system, running throughout the entire body, together with other lymphoid organs and tissues (the spleen, thymus, tonsils and other tissues), provides a structural basis of the immune system and plays a crucial role in body protection [17]. Main functions of the lymphatic system are to provide a return route of the lymph into the blood system and defend the body against infection [18].

The lymphatic system consists of three main parts [17]:

1. a network of lymphatic vessels
2. a fluid inside of the vessels called lymph colourless fluid located between the cells in all body tissues, that contains white blood cells called lymphocytes and circulates throughout the lymphatic system
3. lymph nodes cleanse the flowing lymph

2.2.1 Lymph nodes

Lymph nodes are small, bean-shaped glands composed of lymphatic tissue, widely distributed along the lymphatic routes [19]. Simplified illustration of the lymph node is shown in Figure 2.2. Clusters of lymph nodes nearest to the breast are located in the armpit (called axillary lymph nodes), above the collarbone and in the chest [14]. Axillary lymph nodes provide a majority of the drainage basin for the breast. According to [15], approximately 97 % of the breast lymphatics drain to the axillary lymph nodes, the remaining 3 % drain to the mammary lymph nodes.

Each node is covered by a fibrous capsule that extends inside the tissue a strand called trabecula. The lymph node tissue is differentiated into two distinct regions the cortex, located under the capsule, and the medulla [17]. The most important formations of the cortex and medulla are lymphatic nodules. Each nodule contains lymphocytes, and during an immune response, these nodules develop into centres fighting the infection. Also, a series of lymphatic sinuses, filled with lymph flowing from lymphatic vessels to the nodule, are scattered throughout the node [17, 20].

The primary function of the lymph node is to filter flowing lymph circulating through the lymph vessels all lymph formed in tissues must always pass at least one node before entering back the blood circulation [14, 18, 19]. Lymph is very similar to blood plasma it contains lymphocytes and macrophages

(a) : Illustrative lymph node image (b) : Histopathological lymph node image

Figure 2.2: Detailed illustration of the lymph node anatomy compared to authentic histopathological lymph node image. Illustration taken from [21], histopathological image taken from CAMELYON dataset [22].

cells, but it may also contain microorganisms, waste products and other undesired substances from the tissue [17]. Lymph nodes are responsible for trapping these particles and filtering various pathogens found within the body macrophages and lymphocytes attack and kill them.

Since the lymph nodes play a central role in filtering undesired substances from the cells, it makes them vulnerable to cancer. As was said in Section 2.1, breast cancer typically starts to form in the structure of lobes and ducts [16]. Cancerous cells located in the lobes or ducts start to spread from the tissue via lymph, and they may be trapped in a lymph node, where they start to proliferate. That makes axillary lymph nodes the first place where breast cancer is likely to spread, and recognising metastases in them is one of the most important prognostic factors in breast cancer [14, 19].

2.3 Digital pathology

Digital pathology is a rapidly expanding sub-field of pathology that allows conversion of the classical glass slide, extracted by a pathologist, into a digital image called *whole-slide image* (WSI) that can be uploaded to a computer for viewing and complete electronic management [23]. It represents a fundamental change in the way pathological specimens are viewed. Nowadays, in clinical diagnosis practice, rapid adoption of digital pathology is happening, because manual pathology examination via microscope is time-consuming, tedious and not effective [11, 23, 24]. Compared to that, digital pathology has many