Bachelor Project

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Methods for hierarchical classification of histopathology images

Aleš Král

Supervisor: prof. Dr. Ing. Jan Kybic
Field of study: Open informatics
May 2022

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor prof. Dr. Ing. Jan Kybic, for his great guidance and feedback throughout the making of this thesis. He always promptly responded when I ran into trouble and gave valuable comments and advice.

I wish to express my deepest gratitude towards my family and friends who provided me with encouragement throughout the study years.

Finally, I would like to thank the Czech Technical University in Prague for the education provided.

Declaration

I 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 theses.

In Prague, May 20, 2022

Abstract

In recent years, many methods for hierarchical image classification have been developed showing great promises in the automation of histology image analysis. In this thesis, we implemented some existing methods, mainly the Quadtree approach and the Recurrent visual attention model, for the detection of metastasis in sentinel lymph nodes and compared their performance according to various criteria. We propose modifications to the existing algorithms which improve their ability to correctly distinguish between tumorous and normal tissue.

We also show and utilize different tools and techniques for the preprocessing of whole slide images, namely the patch extraction of the Clustering-constrained attention multiple instance learning method or the Quadtree patch extraction.

The implemented methods have been compared between each other and their references by providing them with test data from the CAMELYON16 challenge. The results were satisfactory for the classification of tiles in whole slide images, with accuracy converging at 87% for the Recurrent visual attention model and 92% for the Quadtree method reaching reference accuracy and sometimes surpassing it. On the other hand, best slide classification from our implementations was around 81% from the Quadtree method and accuracy of the tested CLAM method was around 99% surpassing all other models.

Keywords: machine learning, breast cancer, hierarchical classification, neural network, pathology

Supervisor: prof. Dr. Ing. Jan Kybic
Katedra kybernetiky
FEE CTU
Technická 2,
Praha 6

Abstrakt

V posledních letech bylo vyvinuto mnoho metod pro hierarchickou klasifikaci obrazu, které vykazují uplatn¥ní v oblasti automatizace analýzy histologických obraz·. V této práci jsme implementovali n¥které existující metody, p°edev²ím Quadtree a Recurrent visual attention model, pro detekci metastáz v sentinelových lymfatických uzlinách a porovnali jejich výkonnost na základ¥ r·zných kritérií. Navrhujeme úpravy stávajících algoritm·, které zlep²ují jejich schopnost správn¥ rozli²it mezi nádorovou a normální tkání.

Ukazujeme a vyuºíváme také r·zné nástroje a techniky pro p°edzpracování snímk· celých preparát·, konkrétn¥ extrakci vý°ezk· Clustering-constrained attention multiple instance learning metodou nebo metodou Quadtree.

Implementované metody byly porovnány mezi sebou a svými referencemi tak, ºe jim byla poskytnuta testovací data ze sout¥ºe CAMELYON16. Výsledky byly uspokojivé pro klasifikaci dlaºdic na snímcích celých preparát·, p°i£emº p°esnost konvergovala k 87% u Recurrent visual attention modelu a u metody Quadtree k 92% kde dosahovala referen£ní p°esnosti a n¥kdy ji p°ekonávala. Na druhou stranu nejlep²í klasifikace snímk· celých preparát· z na²ich implementací byla p°ibliºn¥ 81% u metody Quadtree a p°esnost testované metody CLAM byla p°ibliºn¥ 99% která p°ekonala v²echny ostatní modely.

Klí£ová slova: strojové u£ení, rakovina prsu, hierarchická klasifikace, neuronové sít¥, patologie

P°eklad názvu: Metody pro hierarchickou klasifikaci histopatologických obraz·

Contents

Figures

Tables

Abbreviations and acronyms

WSI	Whole Slide Image
RVAM	Recurrent Visual Attention Model
CAMELYON16	CAncer MEtastasis in LYmph nOdes challeNge 2016
AUC	Area Under Curve
GB	Giga Byte
TIF	Tagged Image Format
SVS	Giga Byte
JPEG	Joint Photographic Experts Group
TCGA	The Cancer Genome Atlas
SVM	Support Vector Machines
ResNet	Residual neural Network
H&E	Haematoxylin and Eosin
MIL	Multiple Instance Learning
CLAM	CLustering-constrained Attention Multiple instance learning
ReLU	Rectified Linear Unit
RUMC	Radbound University Medical Center
UMCU	University Medical Center Utrecht
ROC	Receiver Operator Characteristic
AUC-ROC	Area Under the Receiver Operator Characteristic Curve
TPR	True Positive Rate
DAB	DiAminoBenzidine
FPR	False Positive Rate

Chapter 1

Introduction

Breast cancer is one of the main causes of cancer worldwide. Early diagnostics significantly increase the chances of correct treatment and survival.

In the process of histology image analysis for cancer diagnosis, pathologist standardly observes the tissue, its distribution and regularities in cell shapes. Because the tissue slides, called WSIs (Whole Slide Images), are so complex and large in scale (often over 2GB), correct classification of tissue whether it is benign or malignant is often very time-consuming. Fortunately computer-aided analysis has become a rapidly expanding field within the past decade. Computerised scans of stained tissue slides are used by various algorithms to automate tissue classification and aid pathologists.

1.1 Motivation

In the recent past, methods for hierarchical image classification have become quite popular with researchers from different countries tackling on the problem of WSI image classification using various resources and approaches. In practice, these are valuable to automate classification for example of cancer metastasis in lymph nodes of breasts [3] or colon cancer [15]

Most approaches[8][14] try to solve this problem by applying real world practice. WSIs have various staining methods applied to them, most commonly the usage of haematoxylin and eosin (H&E) staining can be observed. This gives the observer, whether it is a human or a machine, information about the most discriminate regions of a tissue slide.

Same as a human analyst, computers have access to different parts of the images at different scales and resolutions. This image decomposition is called the multi-resolution pyramid [11]. It lets researchers use tools for manipulation with these "pyramids" and write algorithms for automatic passing through different levels of magnification and extract information used for image classification.

Many such methods have been developed in recent years, with a difference in performances between each one. Comparison between such methods is crucial, since a researcher or a medical expert could be in search of automatising tissue classification. Having so many options online at our disposal, finding the optimal approach and doing its re-implementation could often be time-consuming.

1.2 Goals

The main focus of this work is to compare different methods for hierarchical image classification, preferably on the same datasets.

In the following chapters some methods for hierarchical image classification will be introduced. Some focus on tissue segmentation to extract only the most informative regions, others try to develop state-of-art deep neural networks which aggregate large quantities of extracted patches to correctly classify given images.

While some methods may be already implemented and used for demonstration [27], others [15][3] will be implemented according to the architecture given to us by the papers. Some articles may not be thorough in the architecture specifications. In these cases, we will try adding modifications, which have improved other models in the past. In the final chapters referenced and modified versions will be discussed for their ability to correctly classify tissue images and how they are able to fulfill their proposed functions.

Chapter 2

Background

2.1 Introduction

Before we begin explaining each method, we first need to get an understanding about tools we will be working with while constructing and testing different approaches. Better understanding of terms explained in this chapter will help us deal with the task.

2.2 Whole slide imaging

Also commonly referred to as "virtual" microscopy [19] involves the scanning of glass slides to produce digital slides. These systems offer pathologists an alternate mechanism to menage and interpret information.

The virtualization is commonly achieved by capturing many small high resolution image tiles and then stitching them together to create a full image for histology analysis. The sequential parts [21] which make the process of digitisation are: image acquisition, storage, editing and display of images.

- Scanners

Whole slide image scanners are specialized devices that are dedicated to acquiring high resolution images of entire slides. These slide scanners consist of 4 main components: light source, slide stage, objective lenses and a high resolution camera for image capture. Whole slide scanners capture images of tissue tile by tile and then "stitch" them to create a digital image of the entire slide.

- WSI

  The slide is standardly captured at high resolutions. Common being x20 or x40 magnification. The resulting structure of the WSI is then called a multi-resolution pyramid.

Figure 2.1: Multi-resolution pyramid showing the same WSI image on different levels of magnification taken from the CAMELYON16 challenge [2]

  Resolution of a WSI is typically expressed in m per pixel. A typical whole slide image scanned at x40 magnification has a resolution of about 0.25 m per pixel with a 24-bit color depth [21].

- Storage

  Acquiring a decently sized dataset leads to huge storage requirements. Since compression of scanned images leads to loss of information and resolution of the image, WSIs are loaded into lossless [25] image formats to preserve data. Lossy data format would be something like JPEG since it compresses the files to a smaller size, but at the cost of data loss, whereas lossless data-formats preserve the information at the cost of a large size. The most common formats used in Medical imaging being .tif or .svs.

2.3 Image tools

Working with gigapixel images is a challenging if not impossible task for standard tools since these are typically designed for images that can be