



Compatible-domain Transfer Learning for Breast cancer classification with limited Annotated Data

Gina cody school of engineering and computer science
Concordia University

Masters in Applied Computer Science

Module: Graduate Seminar Report

Module Supervisor: Dr.Denis Pankratov

Submission Date: 05/08/2022

Speaker: Mohammad Amin Shamshiri

Report By: Vinayak Sareen

Student Id Number: 40186182

Abstract

The study targets the binary classification of malignant breast cancer with limited annotated data. Breast cancer is one of the common kinds of cancer and is the high cause of mortality among women. The classical identification methods require the specialist to review the microscopic images of cytological images which is a tedious and time consuming process. The primary objective of the study is to develop the automated analysis model framework for classification from microscopic cytopathological images in an efficient manner. The model framework can be applied to perform the analysis at scale without need of the specialists. In order to perform classification, there exists various solutions, one of which includes training the convolutional neural networks from scratch which requires fine-tuning hyper-parameters and requires the annotated data in huge quantities. In this research, a framework has been provided to apply the transfer learning based solution from the source histopathological dataset to the target cytological images. The framework relies on transfer learning in which the weights obtained from the pre-training phase are used in the model training with both complete fine-tuning which targets the complete network and updates the weights using the backpropagation throughout the network and partial-fine tuning, which only targets the last few layers as part of the experiments. The proposed framework contains a total eight phases as part of the classification pipeline to provide the mechanism for the binary classification of breast cancers with transfer learning. All the implementation details of the study can be obtained from ¹ footnote link.

¹<https://github.com/ma-shamshiri/Compatible-domain-Transfer-Learning>

List of Abbreviation

1.	ANN	Artificial Neural Network
2.	CNN	Convolutional Neural Network
3.	GPU	Graphical Processing Unit
4.	WHO	World Health Organisation
5.	NHS	National Health Services in United Kingdom
6.	ConvNet	Convolutional Network
7.	ML	Machine Learning
8.	AI	Artificial Intelligence
9.	TL	Transfer Learning
10.	ROI	Region of interest

Contents

1	Introduction	4
1.1	Problem Statement	4
1.2	Motivation	5
1.3	Objectives of the Study	5
2	Technical contribution	6
2.1	Image Segmentation and Preprocessing	6
2.1.1	Color Channel Seperation	7
2.1.2	Image normalization	7
2.1.3	Problems with using the UNET segmentation	8
2.1.4	Intensity Thresholding Segmentation	8
3	Conclusion	10
3.1	Conclusion	10
3.2	Future Works	10

Chapter 1

Introduction

1.1 Problem Statement

The problem of breast cancer among women has been a major concern and leading cause of the mortality rate. According to the WHO(World Health Organization) on the global scale around 502,000 cases of breast cancer among women are reported each year (Jelen, Fevens, and Krzyzak, 2008). Furthermore, on the localized national scale, as per the anticipated reports from the Canadian cancer society, 28,600 women will be detected with malignant breast cancer, representing 25% of the overall new cancer cases detected by the Canadian Cancer Society Society (2020). Furthermore, according to the study conducted in the United Kingdom, the health care agency NHS had reported that 47% trusts do not have the specialized nurses across the country. The lack of specialized medical staff members who can detect breast cancer at the early stage contributes to higher mortality rates (Tan et al., 2017). The classical identification methods require the specialist to review the microscopic images of cytological images which is an inefficient manner even for the trained and qualified medical professionals with appropriate domain knowledge. The problem can be solved using the automation with the CNN and implying the computer vision algorithms to provide the computer aided solution. Convolutional neural networks, which are specific kinds of artificial neural networks, extract data patterns from the images using the convolutional layers and further pass the extracted features to the fully connected layers of the network to learn the features (Wani et al., 2020). However, such networks are highly reliant on the

large volumes of the datasets with appropriate quality which is difficult to obtain in the real world. Moreover, another difficulty with training the ANN is that it is highly reliant on the huge dataset with high accuracy and precision for the training and validation phase, which in real life scenario is difficult to obtain and cost ineffective solution. The convolutional neural networks also require hyper-parameter tuning in order to avoid the overfitting problem in the optimisation step of network training which is a time consuming process and requires domain specialists (Srivastava et al., 2014).

1.2 Motivation

The primary motivation behind the study is to propose a framework based on the transfer learning for binary classification of breast cancer with the limited cytological training samples. The transfer learning framework has eight steps and provides the mechanism to use the pre-trained model's features from the histopathological images to be used with the cytological images. Traditional analysis of the malignancy of the medical images relies on the manual microscopic analysis performed by the specialists. However, such methods can be applied on the large scale with automated processes which will speed up the process for curing breast cancer as early as possible.

1.3 Objectives of the Study

The proposed transfer learning objective of performing superior to the existing solutions in terms of performance with the limited annotated datasamples of cytological images using the domain compatible dataset to avoid high divergence between the source and target models. The proposed framework operates on pre-training the model on the histopathological images and fine tuning the target model for the binary classification of the cytological samples. Although there exists a significant difference in the histopathological image samples and cytological samples, the same domain and knowledge obtained from the pre-trained models using the BreakHis dataset (Spanhol et al., 2015) that contains the 7909 images samples of the breast cancer histopathological with 8 subtypes. The learned model weights are applied to the context of the cytological images.

Chapter 2

Technical contribution

The suggested framework has a total of 8 phases in the classification pipeline which is mentioned in the image presented in table 1.0. The experiments were performed in three scenarios that included pre-training the transfer-learning models on the imageNet, hisbrek dataset and with the random weight initialization in order to evaluate the performance of each of the models with both partial and complete fine-tuning applied to the target models.

2.1 Image Segmentation and Preprocessing

The first phase, targets to obtain the segmented regions of interests from the target source image. In order to perform the image segmentation of the source images of histopathological images, the pre-processing procedure was performed which involves the following sub-procedures which includes color channel separation and image normalization and explained in the detailed manner.

2.1.1 Color Channel Separation

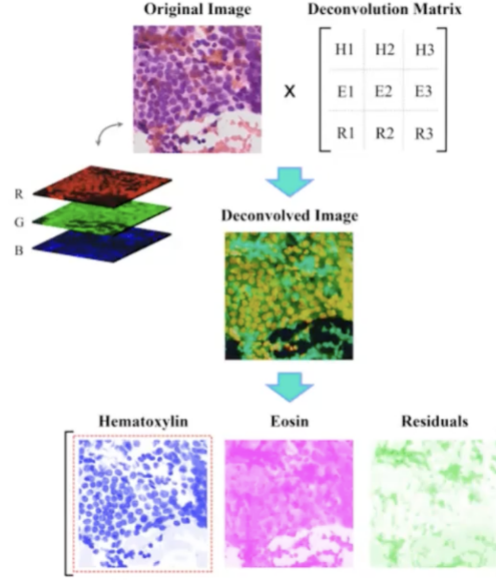


Figure 2.1: Image Credits: MSC Thesis Examination (Presentation) - Mohammad Amin Shamshiri

Color channel separation as shown in the figure 2.1 which aims at decomposing three color channels of the nucleus rather than the other components which are present within the image. In order to perform the color channel separation task, the deconvolutional matrix was applied to extract the feature maps of three distinct kinds that involve hematoxylin, Eosin and residuals. Upon investigation, the hematoxylin feature maps were most prominent and were used for the further normalization and segmentation.

2.1.2 Image normalization

The normalization of the images are performed in order to ensure that the images are easier to process and each pixel in the image is normalized to range within 0 and 1 pixel values rather than 0 and 255, which is easier to train on the convolutional neural networks (Rashid, 2019).

2.1.3 Problems with using the UNET segmentation

In the early phase of the experiments, the semantic segmentation was performed on the hematoxylin images with an objective of grouping the each pixel in the images into the group of nuclei-interior, nuclei-edge or the background using the famous U-Net model, which requires the manual segmentation of the images and that is very time consuming process. The subset of the original dataset was used to evaluate the performance of the segmented images upon splitting the dataset into the training and validation datasets respectively. However, the quantitative analysis of the U-Net segmentation was not possible and can only be evaluated on the visual basis. Therefore as for further investigation the intensity thresholding technique has been opted to deal with such problems and provide a reliable mechanism for image segmentation.

2.1.4 Intensity Thresholding Segmentation

The simple and non-contextual technique focuses on segmenting the images based on the pixel values' replacement with zero in case the corresponding pixel value is less than threshold or 255 otherwise (Wirth, 2004). The result of the segmentation is the binary map of the isolated cell nuclei from other components of the image.

No.	Algorithm	Jaccard Similarity Index			
		Micro	Macro	Weighted	Binary
1	MINIMUM	0.7934	0.7516	0.7994	0.6735
2	LI	0.7834	0.7482	0.7941	0.6728
3	MEAN	0.7899	0.7499	0.8014	0.6654
4	ISODATA	0.7894	0.7464	0.8026	0.6521
5	OTSU	0.7926	0.7448	0.8017	0.6493
6	TRIANGLE	0.6856	0.6538	0.6964	0.5934
7	YEN	0.7645	0.7045	0.7683	0.5843
8	LOCAL	0.6515	0.6193	0.6836	0.5161
9	SAUVOLA	0.5645	0.5506	0.5896	0.4855
	<i>Ground Truth</i>	1.00	1.00	1.00	1.00

Figure 2.2: Image Threshold Segmentation

During the investigation, nine respective algorithms available in the scikit-learn framework as mentioned in the table 2.2 were evaluated. The quantitative analysis of the jacquard index has been opted to compare the performance of various available algorithms which is the ratio of size of the intersection between two sets to size of union of two sets. Upon investigation it was found that the minimum algorithm as mentioned in the table out performs all the other

2.1. IMAGE SEGMENTATION AND PREPROCESSING

algorithms and therefore, has been selected for the segmentation.

Chapter 3

Conclusion

3.1 Conclusion

Based on the experiments performed in the research, it is evident that transfer-learning using the domain compatible histopathological dataset provides high accuracy classification of the cytological image sample. However, applying the partial fine-tuning to the context of the transfer learning is meaningful as compared to using the complete fine-tuning. The study also concludes that employing distinct CNN model architecture does not produce significant differences in the results except for the VGG-19 which performed worse than other models used in the context of transfer learning. The proposed framework yields 6-17% improved accuracy when compared with the existing traditional machine learning solutions and 7% compared to the solutions based on CNN methods. Furthermore, the result of the study eliminates the needs for huge amounts of annotated data for the binary classification of the cytological images while providing the optimal results. The investigation has used a 17 times smaller dataset compared to the existing system while providing 3 per cent optimal results in terms of accuracy.

3.2 Future Works

In order to improve the existing framework, the speaker is looking forward to applying the multi-source domain adaptation techniques which will result in the minimizing the domain-shift between the source and target domain.

3.2. FUTURE WORKS

Furthermore, the utilization of the light-weight network is also taken into consideration in order to diminish the system's dependence. Atlast, the investigation will further be extended to be performed on the image-level scope instead of patch-level to evaluate the systems performance in classification of cancer malignancies in the patients.

Bibliography

- L. Jelen, T. Fevens, and A. Krzyzak, "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies," *International Journal of Applied Mathematics and Computer Science*, vol. 18, no. 1, p. 75, 2008.
- C. C. Society. (2020) Breast cancer statistics. <https://cancer.ca/en/cancer-information/cancer-types/breast/statistics>. Accessed: 2022-07-17.
- Y. Tan, K. Sim, and F. Ting, "Breast cancer detection using convolutional neural networks for mammogram imaging system," in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*. IEEE, 2017, pp. 1–5.
- M. A. Wani, F. A. Bhat, S. Afzal, and A. I. Khan, "Basics of supervised deep learning," in *Advances in Deep Learning*. Springer, 2020, pp. 13–29.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.
- S. Rashid, "What is image normalization?" Oct 2019. [Online]. Available: <https://medium.com/@shoaibrashid/what-is-image-normalization-d8305bf328c0>
- M. A. Wirth, "Lecture notes on image segmentation," February 2004. [Online]. Available: <http://www.cyto.purdue.edu/cdroms/micro2/content/education/wirth04.pdf>