SELF-SUPERVISED REPRESENTATION LEARNING WITH GRAPH NEURAL NETWORKS FOR REGION OF INTEREST ANALYSIS IN BREAST HISTOPATHOLOGY

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Yiğit Özen

December 2020

SELF-SUPERVISED REPRESENTATION LEARNING WITH GRAPH NEURAL NETWORKS FOR REGION OF INTEREST ANALYSIS IN BREAST HISTOPATHOLOGY

By Yiğit Özen

December 2020

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Selim Aksoy(Advisor)

Ramazan Gökberk Cinbiş

Hamdi Dibeklioğlu

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

SELF-SUPERVISED REPRESENTATION LEARNING WITH GRAPH NEURAL NETWORKS FOR REGION OF INTEREST ANALYSIS IN BREAST HISTOPATHOLOGY

Yiğit Özen
M.S. in Computer Engineering
Advisor: Selim Aksoy
December 2020

Deep learning has made a major contribution to histopathology image analysis with representation learning outperforming hand-crafted features. However, two notable challenges remain. The first is the lack of large histopathology datasets. The commonly used setting in deep learning-based approaches is supervised training of deep and wide models using large labeled datasets. Manually labeling histopathology images is a time-consuming operation. Assembling a large public dataset is also proven difficult due to privacy concerns. Second, the clinical practice in histopathology necessitates working with regions of interest of multiple diagnostic classes with arbitrary shapes and sizes. The typical solution to this problem is to aggregate the representations of fixed-sized patches cropped from these regions to obtain region-level representations. However, naive methods cannot sufficiently exploit the rich contextual information in the complex tissue structures. To tackle these two challenges, this thesis proposes a generic method that utilizes graph neural networks, combined with a self-supervised training method using a contrastive loss function. The regions of interest are modeled as graphs where vertices are fixed-sized patches cropped from the region. The proposed method has two stages. The first stage is patch-level representation learning using convolutional neural networks which concentrates on cell-level features. The second stage is region-level representation learning using graph neural networks which can learn the tissue structure. Graph neural networks enable representing arbitrarily-shaped regions as graphs and encoding contextual information through message passing between neighboring patches. Self-supervised contrastive learning improves quality of learned representations without requiring labeled data. We propose using self-supervised learning to train graph neural

networks using vertex dropout augmentation. The experiments using a challenging breast histopathology dataset show that the proposed method achieves better performance than the state-of-the-art in both classification and retrieval tasks.

Keywords: Representation learning, self-supervised learning, graph neural networks, histopathology image classification, content-based histopathology image retrieval, breast histopathology.

ÖZET

MEME HİSTOPATOLOJİSİNDE İLGİ ALANI GÖSTERİMLERİNİN ÇİZGESEL SİNİR AĞLARI İLE KENDİNDEN GÖZETİMLİ ÖĞRENİMİ

Yiğit Özen
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Selim Aksoy
Aralık 2020

Histopatoloji görüntülerinin analizinde derin öğrenme, gösterim öğreniminin elle seçilen özniteliklere göre daha yüksek başarı göstermesiyle birlikte büyük katkıda bulunmuştur. Fakat derin öğrenmenin histopatoloji uygulamalarında iki zorluk etkisini sürdürmektedir. Bu zorluklardan ilki, histopatoloji alanında büyük veri kümelerinin bulunmamasıdır. Diğer alanlarda yaygın olarak kullanılan derin öğrenme tabanlı yöntemler, derin ve geniş modellerin büyük ve işaretli veri kümeleri üzerinde eğitilmesi üzerine kurulmuştur. Histopatoloji görüntülerinin elle işaretlenmesi ise çok zaman alan bir işlemdir. Büyük halka açık veri kümelerinin oluşturulmasında da, tıbbi verilerin mahremiyeti konusundaki çekinceler sebebiyle zorluklar yaşanmaktadır. İkinci zorluk ise, histopatoloji görüntülerinde bulunan ilgi alanlarının değişken boyut ve şekillere sahip ve farklı tanı sınıflarına ait olabilmesidir. Bu zorluğa karşı geliştirilen alışıldık çözüm, ilgi alanlarından kesilerek elde edilen sabit boyuttaki görüntü parçaları üzerinde analiz yapılıp, elde edilen sonuçların ilgi alanı için bir araya getirilmesidir. Bu yöntemler hücre seviyesindeki öznitelikleri çıkarmada başarılı olsalar da, doku seviyesindeki bağlamsal ilişkileri modellemekte etkili olamamışlardır. Bu zorlukların ışığında, çizgesel sinir ağları ve karşılaştırmalı kendinden gözetimli öğrenme yöntemlerini kullanan iki aşamalı bir öğrenme yöntemi önermekteyiz. Bu yöntemde ilgi alanları, düğümleri küçük ve sabit boyutlu görüntü parçalarına karşılık gelen çizgeler olarak modellenir. İlk aşamada, ilgi alanlarından kesilmiş küçük ve sabit boyutlu görüntü parçaları üzerinde, hücre seviyesi niteliklere odaklanan bir evrişimli sinir ağı eğitilir. İkinci aşamada, ilgi alanı seviyesi öznitelik gösterimleri, dokudaki yapıları öğrenebilen bir çizgesel sinir ağı ile öğrenilir. Çizgesel sinir ağları, değişken boyutlu ilgi alanlarının çizgeler olarak modellenmesinin ve bağlamsal özniteliklerin öğrenilmesinin yolunu açmaktadır.

Karşılaştırmalı kendinden gözetimli öğrenme yöntemi ise, etiketli verinin az olduğu durumlarda veriden öğrenilen özniteliklerin kalitesini artırmaktadır. Kendinden gözetimli öğrenme ilk kez rastgele düğüm silme yöntemiyle çizgesel sinir ağlarına uyarlanmıştır. Çalışma kapsamında oluşturduğumuz meme histopatolojisi veri kümesi üzerinde yapmış olduğumuz deneyler, önerilen derin gösterim öğrenimi yönteminin, görüntü sınıflandırma ve görüntü tabanlı içerik erişimi görevlerinde önceki yöntemlerden daha başarılı olduğunu göstermiştir.

Anahtar sözcükler: Gösterim öğrenme, kendinden gözetimli öğrenme, çizgesel sinir ağları, histopatoloji görüntü sınıflandırma, içerik tabanlı görüntü erişimi, meme histopatolojisi.

Acknowledgement

First, I would like to thank my advisor Assoc. Prof. Dr. Selim Aksoy for his patience, kindness, work ethic, and guidance throughout my studies. He set the bar high and helped me to achieve my goals.

I would also like to thank Asst. Prof. Dr. R. Gökberk Cinbiş and Asst. Prof. Dr. Hamdi Dibeklioğlu for accepting my invitation to my defense and for commenting on my thesis.

I thank the expert pathologists Asst. Prof. Dr. Kemal Kösemehmetoğlu, Assoc. Prof. Dr. Sevgen Önder, and Prof. Dr. Ayşegül Üner for making time in their already busy schedules to help establishing the dataset, which made this work possible.

I thank Bulut who has been my companion in this journey, and all my office mates who made our time at the office enjoyable.

Special thanks to my friends Süleyman, Barış, Emrecan, Mehmet, Gökhan Cem, Kubilay, Canol, and Efe. Their long-time friendship is and always will be very valuable to me.

Nothing of this would be possible without the unconditional love and endless support of my mother and father. I am grateful to them for providing me every opportunity they can throughout my life.

I cannot express enough gratitude to my beloved wife Gökçe, who has always believed in me. Her love, unwavering support and encouragement kept me going and gave my work a meaning.

I would also like to thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) for providing financial assistance for my M.S. studies with grant 117E172.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The analysis of disease in histopathology is traditionally made through a microscopic examination of a biopsy by pathologists. The specimen is processed and placed onto glass slides, and tissue is examined under a microscope. In recent years, the advancement in whole slide imaging technology started to digitize the process. Whole slide imaging is an imaging technology that allows glass slides to be scanned at high resolutions to produce digital slides. Digitization of the slides enables the development of more advanced or automated computer aided diagnosis systems in histopathology. With the increasing number of cancer patients, the workload of expert pathologists also increased. Histopathology image analysis aims to serve as an important tool for helping pathologists with the diagnostic process. It can relieve the workload on pathologists and offer more objective analysis of histopathology images.

The first step for histopathology image analysis is modeling image contents by an appropriate representation. This is the fundamental problem for many histopathology image analysis tasks such as cancer classification, content-based histopathology image retrieval, tissue segmentation, and tumor detection. Traditionally, hand-crafted features are extracted from images according to algorithms based on expert knowledge. The most widely used hand-crafted features are Scale-Invariant Feature Transform (SIFT), and Local Binary Patterns (LBP). In

addition to the generic algorithms that can be used for both natural and medical images, there are also hand-crafted features that are designed specifically for histopathology images such as graph-based features like Voronoi diagrams and the distribution of nuclei in the image [1].

Deep learning achieved state-of-the-art results in representation learning, classification, retrieval, detection, and other related tasks. The availability of large training datasets contributed to the success of deep learning. With the dissemination of whole slide imaging technology in recent years, applications of deep learning in histopathology image analysis also started to outperform hand-crafted methods. However, the performance improvement of deep learning-based methods in histopathology domain so far is not as dramatic as in other computer vision domains and their performance do not exceed the pathologists' performance. The difference can be attributed to two major challenges in histopathology domain. The first challenge is the lack of large histopathology datasets. The commonly used setting in successful deep learning application is supervised learning where the data are labeled by expert pathologists. The manual labeling operation in histopathology is expensive and time-consuming. Assembling a large public dataset is also proven difficult due to privacy concerns. A review and approval by an independent ethics committee is usually required to acquire access to any medical data. The second challenge is the necessity for working with multiple regions of interest in while slide images of multiple diagnostic classes with arbitrary shapes and sizes. A region of interest (ROI) is a diagnostically relevant region inside the whole slide image. The typical solution to this problem is to use the bag of features representation where the representations of fixed-sized patches cropped from these regions are aggregated to obtain region-level representations. However, naive methods cannot sufficiently exploit the rich contextual information in the complex tissue structures.

We consider a two-stage learning pipeline for histopathology image analysis. The first stage involves patch-level representation learning where a deep convolutional neural network (CNN) model is trained on fixed-sized patches cropped from the regions of interest. The second stage involves ROI-level representation learning where an aggregation model is trained on ROI images.

The first contribution of this thesis is the proposal to employ a contrastive self-supervised learning method for representation learning in both patch and ROI level. The method tackles the first challenge which is the low data regime. We adapt a self-supervised framework which previously achieved the state-of-the-art results in natural image datasets. Self-supervised learning recently gained attention in computer vision community after the proposed methods started to close the gap between supervised and self-supervised learning [2, 3] after earlier work explored different angles and provided insights to the paradigm [4]. The field of histopathology image analysis can greatly benefit from the self-supervised learning concept due to the low data scenario.

Our second contribution is modeling regions of interest as graphs and employing graph neural networks (GNNs) for patch feature aggregation that can learn contextual information effectively. The graph structure implicitly encodes the spatial relationships across the patches, which can be used to tackle fine-grained representation learning in a holistic manner. Graph neural networks are developed on the generalization of convolution operation found in CNNs [5]. While CNNs work on grid data such as 2D or 3D images and video, GNN convolutions allow modeling irregular structure like the ROI graphs. We conduct experiments on three GNN models that utilize different neighborhood aggregation and pooling operations.

To evaluate the proposed method, we worked with Hacettepe University Pathology Department to construct a new breast histopathology dataset. A large percentage of the specimens that are analyzed in pathology departments are from breast cancer patients, since the disease is the most prevalent form of cancer among women [6]. The dataset contains 1376 regions of interest marked and labeled by expert pathologists on 102 whole slide images collected from 84 patients. Our experimental results show that the proposed method performs better than the state-of-the-art on the breast histopathology dataset.

The remainder of this thesis is organized as follows. The related work involving histopathology image analysis, self-supervised learning, and graph neural

networks is presented in Chapter 2. The dataset description and data preprocessing pipeline is detailed in Chapter 3. The proposed two-stage learning pipeline and representation learning methods are described in Chapter 4. Chapter 5 defines the classification and retrieval problems in histopathology image analysis, describes the experiments which we conducted to measure the effectiveness of the proposed representation learning method, and presents the result of the experiments. Finally, Chapter 6 summarizes the studies in this thesis and the future work.

Chapter 2

Related Work

2.1 Histopathology Image Analysis

Most of the work in histopathology image analysis is focused on the classification problem. Cruz et al. and Spanhol et al. [7, 8] proposed using convolutional neural networks for classification of breast cancer, which outperformed the existing methods based on hand-crafted features. Bayramoglu et al. [9] utilized multi-task learning for a convolutional neural network model with magnification level prediction as the secondary task to boost classification performance. Bejnordi et al. [10] presented a stacked CNN architecture for multi-class breast cancer classification. Mehta et al. [11] extended the U-Net model [12] with the classification task to increase classification accuracy. Das et al. [13] performed classification at multiple magnification levels of histopathology images and combined their results with majority voting.

Some of the above works avoided the challenge of analyzing arbitrarily-shaped regions of interest by only analyzing fixed-sized patches cropped from the regions [8, 9, 13] which prevented exploitation of contextual information. Other works downsized the large region of interest images which causes small-scale biomarkers to vanish [14, 15]. Later, more intelligent aggregation methods were proposed

to better utilize contextual information either by aggregation rules [16, 17] or learned aggregation [18]. More recently, graph neural network architectures that can deal with size and shape variation of region of interest images and encode contextual information via message passing were proposed [19, 20].

Earlier works in histopathology image analysis employed hand-crafted features [21]. Some work combined deep representation learning with histopathology domain knowledge by training the model under some constraints [22]. Transfer learning from models that are trained on natural image datasets to the histopathology image analysis is utilized with and without fine-tuning [23, 24, 25]. Unsupervised neural networks such as autoencoders have also been explored, but their success in large-scale images remain limited [26, 27, 28].

For the content-based histopathology image retrieval task, Vanegas et al. [29] presented one of the first works utilizing deep learning by combining sparse autoencoders for unsupervised feature learning with bag-of-features representation. Zhang et al. [30] proposed hashing based retrieval to time-efficient retrieval of large-scale histopathology images. Ma et al. [31] proposed a region proposal algorithm based on selective search and supervised hashing of regions for retrieval. Zheng et al. [32] presented a two-stage method which involves feature extraction from superpixels using a sparse autoencoder and supervised hashing for retrieval. In another work, they handled the retrieval of variable size regions by unsupervised hashing of patches and predetermined distance measures for comparing regions consisting of hashed patches [33]. Most recently, label-based supervised training of graph neural networks on regions of interest and using graph embeddings for retrieval is proposed [34].

2.2 Graph Neural Networks

Although first applications of neural networks to graphs based on recurrent architectures appeared earlier [35], the GNNs that are similar to CNNs with generalized convolutions were presented recently [36, 37, 38]. Kipf et al. [39] used

stacked graph convolutions for semi-supervised classification. Morris et al. [40] showed that previous GNN architectures have the same expressiveness as the one-dimensional Weisfeiler-Leman graph isomorphism heuristic, and proposed a k-dimensional GNN architecture that can take higher-order graph structures at multiple scales into account. Ying et al. [41] introduced a hierarchical representation learning method to graph neural networks.

Applications of graph neural networks in histopathology image analysis is so far limited. Zheng et al. [19] proposed using GNNs that are trained using class labels as feature extractors for histopathology image retrieval. Recently, Wang et al. [42] presented performance improvements with using graph convolutional networks over CNN-based features in weakly-supervised prostate histopathology classification. In another medical domain, Li et al. [43] utilized GNNs to identify biomarkers of autism spectrum disorders in brain networks contructed by fMRI.

The existing work involve supervised or weakly-supervised training of GNNs. None of the existing work consider self-supervised training of GNNs. The studies that involve the application of GNNs in the medical domain compare it with pooling-based or multiple instance learning methods. There is a lack of studies that compare the effects of the GNN architectures on the histopathology image representation learning.

2.3 Self-Supervised Learning

Self-supervised learning methods involve training a model for a proxy task which forces the model to learn to extract semantic features that also benefits the aimed downstream task. Examples of proxy task that were used in earlier works are inpainting [44], colorization [45], and puzzle solving [46]. A method of self-supervised learning based on contrastive predictive coding where future outputs of the encoder model are predicted by an autoregressive model achieved the state-of-the-art performance at the time [47]. Later, Chen et al. [48] proposed contrastive learning on top of stochastic augmentations which outperformed earlier

approaches with a simple architecture. Momentum Contrast (MoCo) [49, 50] is another contrastive self-supervised learning method in which a second, slowly progressing encoder is used to build a dictionary for the contrastive loss. Recently, self-supervised learning methods are employed for feature extraction from small histopathology image patches [2, 51].

Although the studies on self-supervised methods show promising results which would be especially beneficial in histopathology setting where labels are scarce, there is a lack of comparative studies on the subject.

Chapter 3

Dataset

3.1 Dataset Description

We constructed a new breast histopathology dataset using 102 whole slide images (WSI) that were digitized from specimens collected from 84 patients. The haematoxylin and eosin (H&E) stained specimens were selected from the archives of the Department of Pathology at Hacettepe University based on their slide-level diagnoses. H&E is the most widely used stain in histopathology. The hematoxylin stains nuclei blue, and eosin stains the cytoplasm pink. Other structures take on different combinations of these colors. The WSIs were acquired at 40× magnification by using an Olympus slide scanner. The resulting average image size was 170,000 × 132,000 pixels. 1,376 regions of interest (ROI) were annotated by experienced pathologists in free form with no shape and size restrictions. The resulting annotations were collected into 4 diagnostic classes: *benign* (including samples containing non-proliferative changes, apocrine metaplasia, usual ductal hyperplasia, columnar cell hyperplasia, flat epithelial hyperplasia, and intraductal papilloma without atypia), *atypia* (including samples containing atypical ductal hyperplasia, atypical lobular hyperplasia, and intraductal papilloma with atypia), *in situ* carcinoma (including both ductal carcinoma in situ and lobular carcinoma in situ), and *invasive* carcinoma. The per-class ROI size statistics are shown in

Table 3.1: ROI size statistics per diagnostic class in number of pixels at 10× magnification. Rows show the average ROI size, the standard deviation of ROI sizes, and the ratio of the largest ROI size to the smallest one, respectively.

	Benign	Atypia	In Situ	Invasive
Average	$1308K$	$473K$	$2815K$	$12568K$
Standard deviation	$2510K$	$711K$	$4948K$	$17822K$
Max-min ratio	977.2	210.8	941.1	762.5

Table 3.1.

3.2 Preprocessing

Since the specimens were stained at different times following different procedures, there is a great variation in their color distributions. To eliminate the effects of staining differences in model learning, we performed stain normalization as a preprocessing step. We first applied color deconvolution [52] to estimate the stain matrix of each slide. The method involves estimating the pure haematoxylin and eosin colors in red-green-blue (RGB) color space and derive the RGB-to-H&E conversion matrix. To make the estimation more robust, haematoxylin stain vector estimation considered only the pixels inside nuclei masks which were automatically generated using a pre-trained convolutional network, and eosin stain vector estimation considered the remaining regions excluding high luminosity regions which correspond to background. Then, the estimated matrix is used to convert 3-channel RGB images to 2-channel H&E images. Then, the histograms of haematoxylin and eosin channels of each slide are matched to a target slide chosen from the dataset [53]. Finally, the inverse conversion is performed to return to the RGB color space.

We partitioned the dataset into four folds by using ROI-level diagnosis labels. The split is constrained to make all WSIs and ROIs from the same patient fall under the same fold. We employed a genetic algorithm with a fitness function based on class distributions to find a good split that achieves similar slide-level and ROI-level class distributions among the folds while adhering to the constraint.

Table 3.2: Class distribution of slides and ROIs in training, validation, and test sets. Note that a slide can contain multiple ROIs corresponding to different diagnostic labels, resulting in a multi-label setting for each slide. Thus, the numbers of slides for each diagnostic class in the table do not sum up to the total number of slides for a given set.

		Benign	Atypia	In Situ	Invasive	Total
Slide	Training Set	42	20	22	15	53
	Validation Set	20	10	11	7	26
	Test Set	18	11	10	7	23
	Total	80	41	43	29	102
ROI	Training Set	291	73	192	118	674
	Validation Set	144	35	95	56	330
	Test Set	147	38	128	59	372
	Total	582	146	415	233	1376

We use two folds to train the models, one fold as the validation set for model selection, and one fold as the test set for performance estimation. The folds are assigned to the subset randomly once, and the same assignment is used in all experiments. The resulting ROI-level and slide-level class distributions of the three sets are given in Table 3.2.

Chapter 4

Representation Learning

This chapter proposes a new method for learning semantic visual representations for variable and arbitrarily-sized regions of interest in whole slide histopathology images. Learning efficient and generalizable visual representations of tissue images is critical for the performance of downstream tasks such as cancer type classification and image retrieval for computer-assisted diagnosis. The proposed method can work without any labeled data, or can utilize the unlabeled data with labeled data which is generally the case in histopathology datasets due to the annotation costs. The method is compared with other histopathology image representation learning methods that are the bases for many histopathology image analysis methods in the literature.

4.1 Two-Stage Learning Pipeline

Convolutional neural networks (CNNs), after their initial success in natural images domain [54], were also successfully adapted to medical image analysis [55]. In digital pathology, the sizes and shapes of the diagnostically relevant regions in whole slide images differ greatly. The de facto method for handling variable size input in deep learning-based visual representation learning that have been the state-of-the-art for almost a decade in other computer vision domains is padding and resizing all images to a fixed size. Due to the great maximum-over-minimum size ratio in histopathology region of interest images, cell-level markers that are too small relative to the target image size vanish in downsampling or padding operations. Furthermore, resizing prevents utilizing size-based markers. Earlier works have overcome this problem by splitting the pipeline into two stages. In the first stage, the image is divided into fixed-size patches, and features are extracted from the patches. In the second stage, the patch features are aggregated into ROI-level features.

The proposal of this thesis improves the pipeline in two ways. First, it models the arbitrarily-shaped regions of interest as graphs and employs graph neural networks in the aggregation stage. Second, it uses self-supervised learning which allows utilization of unlabeled data and improves the quality of learned representations. With a careful choice of patch sampling and graph construction methods, the first stage models the small-scale cell features of the tissue, while the second stage models large-scale structures in tissues such as ductolobular system in breast tissue.

The proposed learning pipeline is given in Figure 4.1. In the first stage, fixed-sized patches are sampled from regions of interest, and used to train a patch-level convolutional neural network in a self-supervised fashion. In the second stage, regions of interest graphs are constructed with sampled patches. The trained model from the first stage is used as patch-level feature extractor. The ROI-level graph neural network model is also trained in a self-supervised fashion. In the case of existence of ROI labels, the ROI-level model can be fine-tuned by

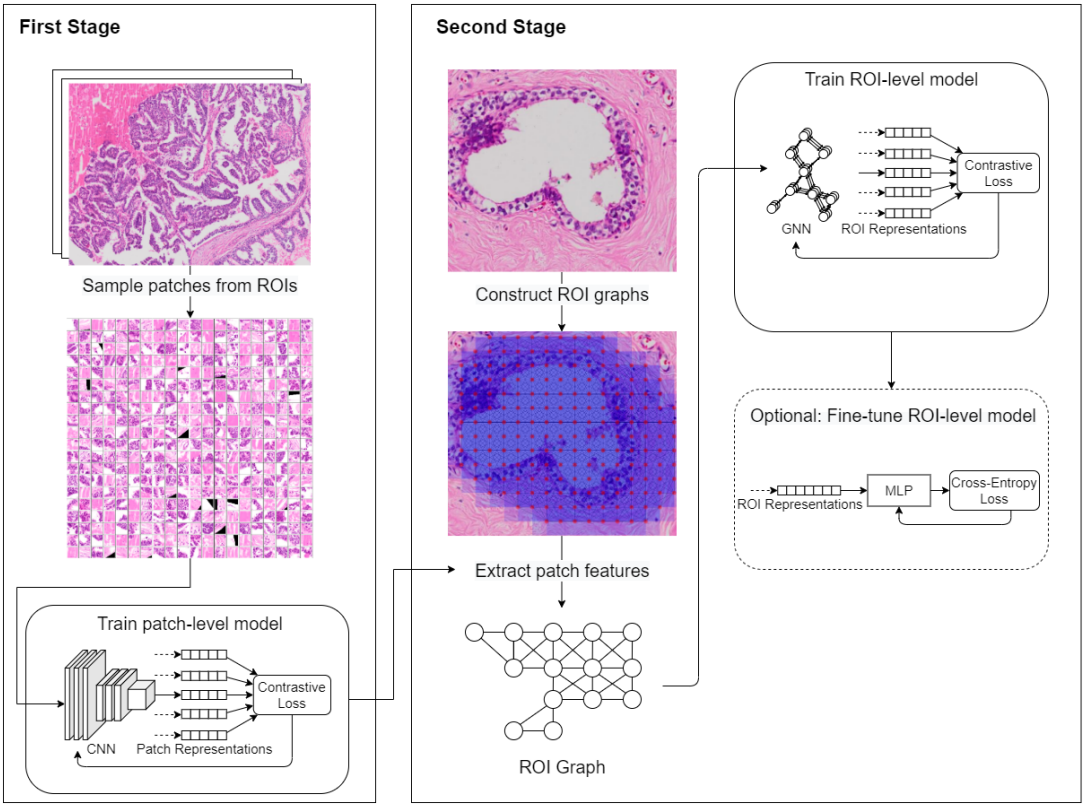

Figure 4.1: The proposed two-stage learning pipeline. In the first stage, a patch-level feature extractor is trained on the patches that are cropped from the ROIs using the contrastive self-supervised method. In the second stage, ROIs are modeled as graphs where vertices are fixed-sized patches. The frozen patch-level model from the first stage is used to extract patch features. Then, an ROI-level model is trained on ROI graphs using the contrastive self-supervised method. Optionally, the ROI-level model can be fine-tuned with the cross-entropy loss using ROI labels.

supervised training.

4.1.1 Patch Sampling

Patches are obtained from regions of interest using sliding-window technique of which the patch size and step size are the two parameters. All patches that fall into a region are used without an extra condition unlike other approaches that discard some of the patches using a fixed algorithm, such as calculating nuclei

density and applying a threshold. This method avoids additional assumptions associated with adding an auxiliary signal that tells the model where to look. Also, including white spaces can help the ROI-level model to better identify ducts and lobes. This thesis argues that the fixed patch size relative to the magnification level should be selected so that the tissue content falls inside a patch contains similar cells in terms of diagnostically important small-scale features such as nuclei shape and presence of mitosis; but does not contain large-scale features such as the shapes of the walls of the ducts.

4.1.2 Region of Interest Graphs

A region of interest graph is a representation of a region of interest by an undirected graph whose vertices correspond to the fixed-size patches and edges correspond to spatial proximity relation between the patches. The vertices of the graph is obtained using the method described in the previous section. An edge between two vertices is inserted to the ROI graph if the distance between the centers of two patches corresponding to the vertices is smaller than a threshold parameter.

4.1.3 Self-Supervised Learning Framework

Self-supervised learning only recently gained attention in computer vision community after the proposed methods started to close the gap between supervised and self-supervised learning [2, 3] after earlier work explored different angles and provided insights to the paradigm [56, 57]. Medical image analysis domain quickly followed the trend and several studies that employ self-supervised methods for learning representations have published recently [58].

We adopt the SimCLR framework [48] for learning representations without labeled data using a contrastive loss function. SimCLR has achieved state-of-the-art results on the ImageNet dataset with an architecture simpler than its

alternatives such as Contrastive Predictive Coding [47].

SimCLR framework comprises a stochastic data augmentation module, a neural network encoder, a neural network projection head, and the normalized temperature-scaled cross entropy loss as the contrastive loss function. The data augmentation module takes as input a data sample, and outputs two different augmented views of this example. The encoder extracts feature vectors from the augmented data points. Finally, the projection head maps the feature vectors to the space where contrastive prediction loss is calculated.

During the end-to-end training of the model, a random mini-batch of size M is fed to the augmentation module, resulting in $2M$ data points. Given the positive pair of a data point which originated from the same example, the other $2(M-1)$ data points are treated as negatives. The loss function between a positive pair of data points (i, j) is defined as

$$l_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\Sigma_{k=1}^{2M} \mathbb{I}_{[k \neq i]} \exp(sim(z_i, z_k))/\tau} \tag{4.1}$$

where $\mathbb{I}$ is the 0-1 indicator function, z_i, z_j are the outputs of the prediction head, sim is the cosine similarity function, and τ is the adjustable temperature parameter. The loss function in (4.1) is computed across all positive pairs in a mini-batch:

$$L = \sum_{(i,j)|s(i)=s(j)} l_{i,j} \tag{4.2}$$

where $s(.)$ indicates the source image of an augmented image.

The design of the stochastic data augmentation module that uses a composition of several data augmentation operations is vital for learning effective representations. The design should be determined considering the target domain. This thesis proposes different augmentation modules for patch-level and ROI-level learning stages which are detailed in Section 4.2.3 and 4.3.3.

A multi-layer perceptron with one hidden layer is used as a projection head in both stages. Previous studies show that a learnable non-linear transformation

between the representations and the contrastive loss function results in a better performance than linear or no transformation [59].

Contrastive learning benefits from larger batch sizes and long training. Using a small patch size in the first stage makes using large batch sizes and fast training possible without large compute clusters.

4.2 Patch-Level Representation Learning

The section compares visual representation learning methods for patch-level learning stage. Due to the annotation costs, patch-level labels are generally missing in histopathology datasets. Earlier studies worked in two machine learning settings to learn representations for patch data. The first is transfer learning which can be defined as exploiting knowledge gained while learning one task to another task. The second is weakly-supervised learning where inaccurate or noisy labels are used as supervision signal. This thesis proposes applying self-supervised learning which utilizes the data itself as the supervision signal in the lack of labels.

4.2.1 Transfer Learning

Recent successes of convolutional neural networks (CNNs) achieved state-of-the-art performance in image classification applications [60, 61, 62]. However, training them with random initializations requires a large amount of data to achieve high performance. This makes them challenging to use in domains like medical image analysis where generating and annotating data is expensive. Transfer learning was proposed to address this shortcoming [63]. A common transfer learning method is to train a model on a large source dataset, and then fine-tune the pre-trained model on the target dataset. During fine-tuning, the model starts with the knowledge to extract relevant features from the source domain. Transfer learning can improve the performance when retaining some of this knowledge helps with the task.

In the histopathology image analysis setting, patch-level annotation or labels are generally not available. Transfer learning in this case is unsupervised, i.e. the supervised fine-tuning step is skipped. Due to the lack of large public histopathology datasets, earlier works used models pre-trained on ImageNet [23] dataset. Since features representing shapes, colors and textures in cell-level analysis is still relevant in the target domain, the method, despite of its limits, achieved some success [24, 25]. This thesis applies the method of using an ImageNet-pretrained CNN as a feature extractor on the breast histopatology dataset to compare unsupervised transfer learning approach with the alternatives.

4.2.2 Weakly-Supervised Learning

Weakly-supervised learning is an umbrella term covering several types of learning settings where a machine learning model is trained with weak supervision [64]. In the patch-level representation learning case, where the ground truth is unknown in patch-level but known in ROI-level, using ROI labels is called inaccurate supervision. The reason is that patch-level labels are not always consistent with the ROI-level label. The mapping from patch content to ROI classes is one-to-many, that is certain patch content can occur in one or more types of ROI since ROIs contain a mixture of cells and structures. For example, a patch containing cancerous ductal cells spreading into surrounding breast tissue may only occur in ROIs within invasive carcinoma class, but a patch with only stroma tissue can occur in any class. Therefore, the corresponding ROI label for a patch is an inaccurate label. Yet, earlier studies used weak supervision to train cancer type classifiers [8]. More recent studies use weak supervision for patch-level learning, coupled with ROI-level supervised training [65]. This thesis also adopts this strategy as the first stage in the pipeline described in Section 4.1 to compare it with the other patch-level representation learning methods. In this setting, the patch-level model is trained in a supervised fashion with the cross-entropy loss using the label of the ROI from which the patch is cropped as the patch label.

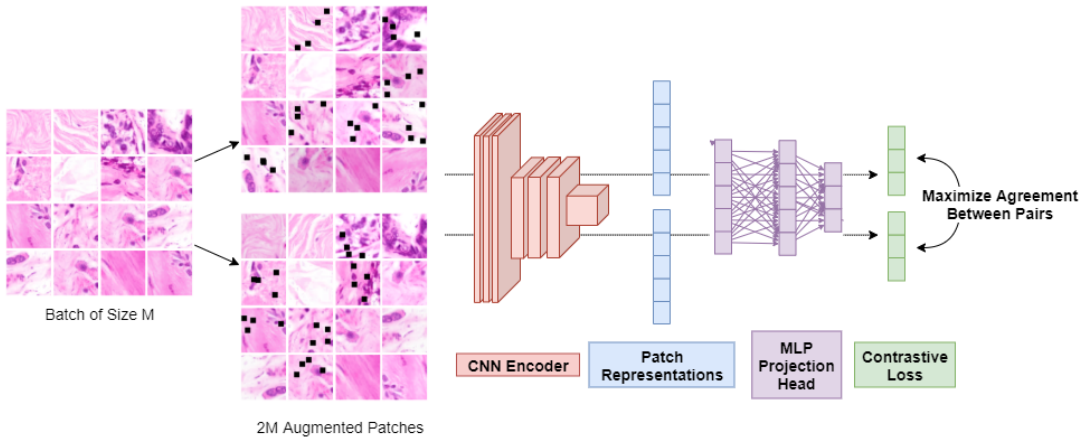

Figure 4.2: Contrastive self-supervised learning of patch-level representations. The patches in the batch are randomly augmented twice. The resulting pairs of patches look different, but the diagnostically relevant information is preserved. The CNN encoder and MLP head are trained end-to-end using the contrastive loss function which is calculated using the transformed representations shown in green. After training, the MLP head is removed, and the CNN encoder is used to extract patch representations shown in blue.

4.2.3 Self-Supervised Learning

This thesis proposes using self-supervised learning framework described in Section 4.1.3 for patch-level representation learning. The process for training patch-level model using SimCLR framework is described in Figure 4.2. The patches are sampled from regions of interest as described in Section 4.1.1. The stochastic data augmentation module is designed to the contrastive prediction task harder while refraining from mixing diagnostically relevant visual clues. Size, shape, and contrast level of cells are such clues. A random part of the patch is cropped and rescaled to the original size. The crop size itself is random, but not smaller than 95% of the original size. The patch is randomly flipped and rotated, blurred, changed brightness, and several regions are cut out [66]. A CNN is used to extract feature representation from the augmented patches. A multi-layer perceptron (MLP) with one hidden layer is used as to non-linearly project the representations to a latent space where NT-Xent loss given in Equation (4.1) is applied to train the CNN and MLP.

4.3 ROI-Level Representation Learning

Going from patch-level representation to ROI-level representations requires an aggregation method. The most simplistic approach is maximum or average pooling over feature representations of all patches which comprise a region of interest [8, 67]. More recent works assume presence of a hidden variable associated with each patch indicating the patch's contribution to the ROI-level discrimination, and propose methods to learn that hidden variable [17]. This thesis proposes using graph neural networks to enable learning context through relationships between neighboring patches to encode the large-scale structural information.

4.3.1 Multiple Instance Learning

Aggregation of patches can be modeled using multiple instance learning (MIL) bag of instances paradigm where patches are the instances and ROIs are the labeled bags. The standard multiple instance (SMI) assumption states that, in a binary classification problem, a bag is positive if there exists at least one positive instance in the bag. SMI assumption can be modeled using max-pooling over feature representations of all patches which comprise a region of interest. Using this method for ROI-level representation learning requires multi-class classification problem to be adjusted by ordering the classes to allow maximum operation over them. In cancer classification task, severity of the disease is used to order diagnostic classes. However, presence of a highly-severe type of instance does not necessarily mean highly-severe type of bag. The complex subtypes of the main diagnostic classes greatly affect the prognosis of the disease as the progression is not always linear. Max-pooling in this multi-class classification task fails to model these details.

A better suited assumption is threshold-based multiple instance (TMI) assumption which generalizes the SMI assumption. TMI assumption states that the number of instances that belong to a type which a bag contains must be

greater than a lower bound to be labeled. It can be modeled using average pooling or voting.

A further generalization called count-based multiple instance (CMI) assumption enforces both lower and upper bound for the number of instances. Other studies proposed a method called decision fusion that pools the distribution of predicted class probabilities in patch-level [18]. Note that this method cannot be combined with the unsupervised transfer learning and self-supervised learning methods described in Section 4.2.1 and 4.2.3. Another method is attention-based aggregation where the ROI representation is the weighted average of the feature representations [68]. Let $R = P_1, ..., P_K$ be an ROI bag of P patches, the ROI representation G is defined as:

$$G = \sum_{k=1}^{K} a_k P_k \tag{4.3}$$

where:

$$a_k = \frac{exp\{w^T tanh(VP_k^T)\}}{\sum_{j=1}^{K} exp\{w^T tanh(VP_j^T)\}} \tag{4.4}$$

where $w \in \mathbb{R}^{L\times 1}$ and $V \in \mathbb{R}^{L\times M}$ are learned parameters and $tanh$ is the hyperbolic tangent function for non-linearity. The learned weights enable the model to attend to the patches that are relevant for the bag-level discrimination.

In our experiments, we compare average pooling and attention-based MIL methods to the proposed GNN-based method for aggregation.

4.3.2 Graph Neural Networks (GNNs)

The graph representation of a region of interest defined in Section 4.1.2 can be modeled using Graph Neural Networks (GNNs). Motivated by CNNs in deep learning, GNNs are developed on new generalizations of convolutions. While the standard convolutions operate on grid data, the generalized convolution operation

can work on the irregular structure of graphs with variable number of vertices and edges [69].

Three common building blocks of GNN models are graph convolutions, local pooling, and global pooling [5]. Graph convolutions enable encoding contextual information by passing vertex features through graph edges followed by a non-linear activation. Local pooling makes the learned representations hierarchical similarly to the maximum and average pooling operators in convolutional neural networks. Global pooling is similar to local pooling, but is applied at the end of the network to aggregate vertex representations into a graph representation.

In its general form, a graph convolution operator can be defined as

$$X' = \Phi(A, X; \Theta) \tag{4.5}$$

where $X \in \mathbb{R}^{N \times C}$ is the input feature matrix of N vertices with C input channels, A is the adjacency matrix, Θ is the set of trainable parameters, and $X' \in \mathbb{R}^{N \times F}$ is the output feature matrix with F channels.

We consider two convolution operators. The first one is proposed by Kipf et al. [39]. The convolution operation for a single vertex can be defined as

$$X'_i = \Theta \sum_{j \in N(i)} \frac{1}{\sqrt{d_j d_i}} X_j \tag{4.6}$$

where X_i is the feature vector of vertex i, Θ is a learned parameter, $N(i)$ is the set of neighbors of vertex i, and d_i is the degree of vertex i.

The second is proposed by Morris et al. [40] and defined as:

$$X'_i = \Theta_1 X_i + \sum_{j \in N(i)} \Theta_2 X_j \tag{4.7}$$

For local pooling, we consider top-k pooling operator from [70, 71, 72] to compute pooled graph (X', A') using

$$y = \frac{Xp}{\|p\|} \tag{4.8}$$

$$i = top_k(y, k) \tag{4.9}$$

$$X' = (X \odot tanh(y))_i \tag{4.10}$$

$$A' = A_{i,i} \tag{4.11}$$

where p is the learned parameter for projection, $\|.\|$ is the L_2 norm, top_k selects the top k indices from a given input, $\odot$ is element-wise multiplication, and $._i$ selects the elements at the specified indices.

For global pooling, vertex features can be averaged across the vertex dimension to produce a single graph feature vector G as

$$G = \frac{1}{N} \sum_{i=1}^{N} X'_i \tag{4.12}$$

Like other deep learning models, the building blocks of graph neural networks can be combined in different ways. There are many more graph convolution and pooling operators than we describe above which can be used to construct a GNN. In this study, we use the two aforementioned convolutions, top-k pooling and global vertex averaging to construct and conduct experiments on the GNN variants.

4.3.3 Self-Supervised Learning with GNNs

Unsupervised learning of GNNs were proposed in computer vision literature using autoencoder architectures where the encoder embeds the graph into a latent representation which decoder uses to reconstruct the graph [69]. In medical image analysis, application of GNNs remains limited to supervised learning [42, 43].

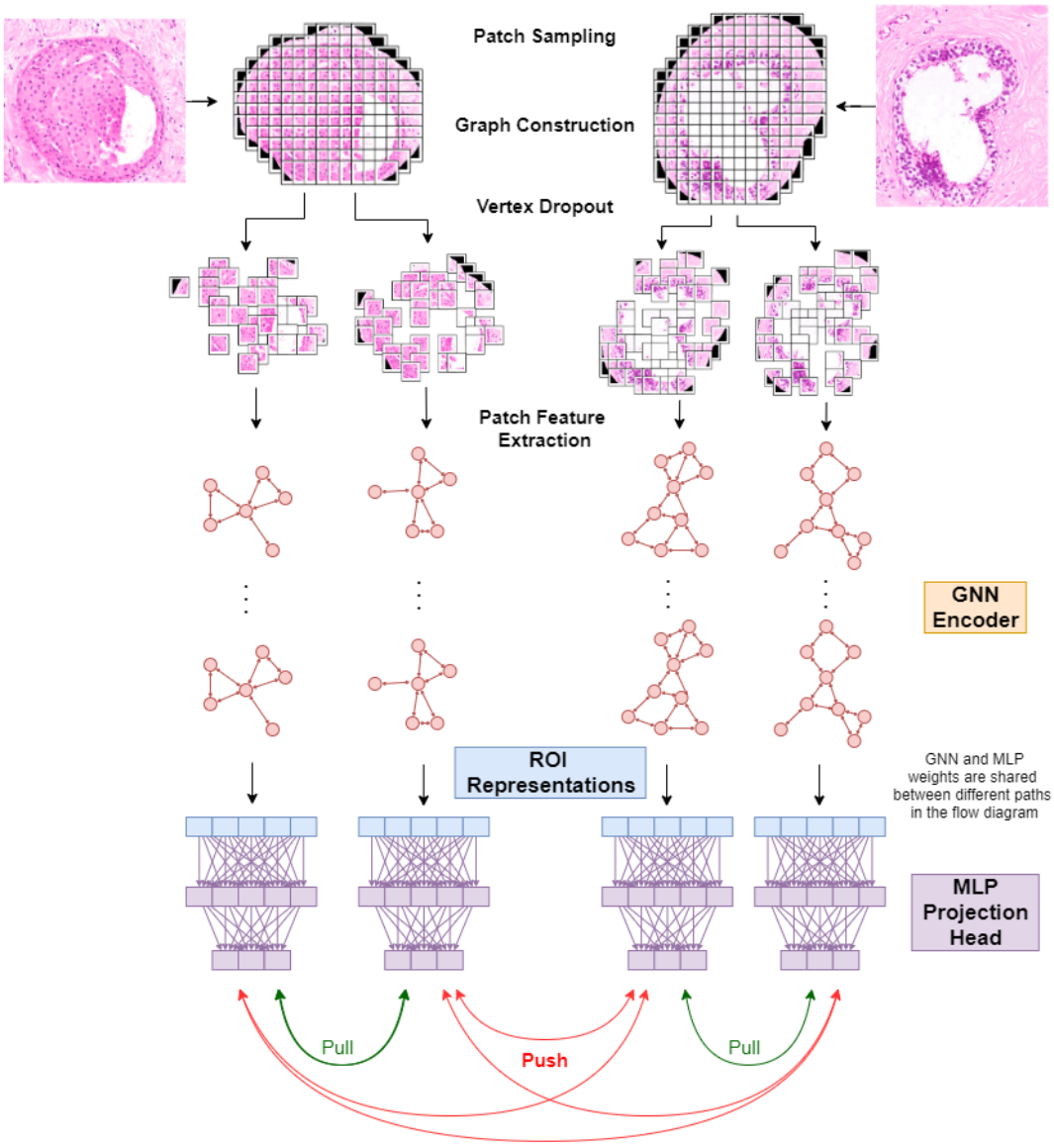

Figure 4.3: Contrastive self-supervised learning of ROI-level representations. Only two regions of interest from a batch are shown. We sample patches from the regions, and construct their graphs. The stochastic augmentation module creates two pairs of graphs. We used a two-layer MLP, shown in purple, for transforming learned representations before calculating the loss. The contrastive loss function causes the model output similar representations for augmentations of a region, and different representations for different regions. The network is trained end-to-end. After training, we can remove the MLP, and use GNN shown in yellow to calculate representations of regions for retrieval operations.

This thesis proposes a new method to train GNNs in a self-supervised fashion without annotated data.

The effectiveness of the contrastive self-supervised learning method described in Section 4.1.3 depends on the augmentations performed on the input data. We propose randomly removing, or dropping out, a portion of vertices from the graph as augmentation. With the careful choice of patch size and dropout ratio, diagnosis from the augmented region of interest graph can still be performed by an expert pathologist. By training using contrastive loss on the augmented graphs the GNN models can acquire this knowledge. Furthermore, we show that vertex dropout alone can be a strong enough augmentation, allowing patch-level feature extraction to be performed once for each graph before training which greatly reduces the model training costs. The process for the proposed self-supervised training of GNNs on region of interest graphs is shown in Figure 4.3. This process allows employment of any GNN architecture.

Chapter 5

Experiments and Results

This chapter covers two downstream tasks for region of interest analysis: classification and retrieval. For both tasks, we conduct experiments using the dataset presented in Chapter 3 and region of interest representations learned with the methods presented in Chapter 4. We compare the effectiveness of learned representations with respect to the performance metrics for the given tasks. In the end, we show that the proposed representation learning method achieved better performance than the alternatives in both classification and retrieval tasks on the breast histopathology dataset.

5.1 Classification

Automated classification of histopathology images is one of the most popular problems in medical image analysis [73]. Multi-class classification of gigapixel histopathology images with only small labeled datasets available for machine learning is still a challenge. Learning effective representations without label-based supervision is greatly beneficial for the histopathology image classification task. Classification for the breast histopathology dataset is defined as the problem of assigning one of the four categories to a given region of interest image.

5.1.1 Evaluation Method

The experiments are implemented in Python language using PyTorch [74] machine learning framework and Lightning [75] research framework. The hyper-parameters are optimized using random search and validation set performance. All training tasks are carried out using a single machine with two NVIDIA RTX 2080Ti GPUs.

Quantitative analysis is performed by comparing the class predictions of the models to the labels assigned to the regions of interest by the pathologists. We apply the three-way holdout method for model selection and performance estimation using the training, validation, and test sets described in Chapter 3. Given a test set $\{(X_i, Y_i)\}_{i=1}^{N}$ where X_i denotes the i'th region of interest and Y_i its reference label, let $Y_i' = f(X_i)$ denote the label predicted by a classifier. The accuracy of the classifier on the test set is defined as $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[Y_i = Y_i']$ where $\mathbb{I}$ is the 0-1 indicator function. The balanced accuracy is defined as the average of per-class accuracy metrics.

5.1.2 Patch-Level Representation Learning

We compare three methods for patch-level representation learning to evaluate its effect on the classification performance:

- Transfer learning using an ImageNet-trained model
- Weakly-supervised learning using ROI-level labels
- Self-supervised learning using a contrastive loss

5.1.2.1 Model Configurations and Training

Patches with size 64×64 pixels are extracted from regions of interest in $10\times$ magnification level using sliding-window with 32 pixel steps resulting in 50%

overlap between consecutive patches. The maximum patch count that can be extracted from an ROI is limited to 20,000 to balance the ROI distribution of patches, while the largest ROI has 2 billion pixels, and the smallest ROI has 116,000 pixels. If the number of patches sliding-window generated exceeds the threshold, 20,000 patches are randomly selected.

For all three models, the same EfficientNet-B0 [61] CNN architecture is used as encoder. EfficientNet is a family of image classification models that achieved better classification performance than earlier models such as ResNet [60] while being more efficient in terms of time and memory resources. B0 is the smallest member of the family with 5.3 million parameters. We employed this model due to GPU memory limitation. Other works show that self-supervised learning methods like ours greatly benefit from larger models [59].

For the transfer learning setup, the parameters of the model are transferred from an ImageNet-trained model which is downloaded from an online model zoo [76]. No further training is performed.

For weakly-supervised learning, the final fully connected layer of the model is replaced to output four logits for four classes. Parameters of the remaining layers are transferred from the same model used in the transfer learning setup. The model is optimized using cross-entropy loss with a batch size of 256. Adam optimization algorithm with corrected weight decay regularization [77] is used with 1×10^{-6} weight decay factor. One-cycle learning rate scheduling policy [78] is employed with 20000 iterations warm up and 1×10^{-3} maximum learning rate.

For self-supervised learning, the final fully connected layer of EfficientNet is removed. A 2-layer MLP projection head with a hidden dimension of 512 and output dimension of 128 is added. All parameters are initialized randomly using Kaiming-uniform method [79]. The encoder and the projection head are optimized end-to-end using NT-Xent loss defined in Equation (4.1) with a batch size of 1024 which results in 2048 patches after the stochastic augmentation step. The loss temperature hyperparameter is set to 0.1. Adam optimization algorithm with corrected weight decay regularization is used with 1×10^{-6} weight decay factor.

Cosine annealing learning rate scheduling is utilized with linear warm up of 10000 iterations and 1×10^{-3} maximum learning rate. After training is completed, the projection head is removed.

During the training of both models, the classes are oversampled inversely proportional to their example counts to counter the class imbalance in the dataset. Balanced accuracy, i.e. the macro average of per-class recalls, is computed on the validation set at the end of every epoch. The models are trained for 100 epochs, and the model in the epoch with the best balanced accuracy is selected. We observed that training would have stopped before the 100th epoch if we used early stopping based on validation loss, accuracy, or balanced accuracy.

The trained patch-level models are used as feature extractors for the supervised ROI classification task. ROI classification task is repeated with average pooling aggregation and GNN-based aggregation. Average pooling aggregation method computes the average of feature representations of all patches which comprise a region of interest. Then, a two-layer MLP which outputs four class scores is trained with cross-entropy loss function. GNN-based model consists of stacked graph convolutions given in Equation 4.6 which outputs four class scores is trained with cross-entropy loss function.

5.1.2.2 Results

The results of the experiments which compare patch-level learning methods are given in Table 5.1 and 5.2. The results show that the weakly-supervised method performed the best when the labels of all examples in the dataset are utilized in training. The proposed self-supervised method which does not require labels tied with the weakly-supervised method with the number of labeled examples reduced to 10% when combined with average pooling aggregation, and outperformed it when combined with the GNN-based aggregation. We argue that when the patch representations correspond to the cell-level features, simple averaging of those features do not discriminate ROI-level diagnostic classes. The comparison in Table 5.1 where the balanced accuracy is similar for each method, but overall

Table 5.1: ROI classification performances for different patch-level representation learning methods followed by average pooling aggregation.

Method	Label fraction	Accuracy	Bal. Accuracy
Transfer Learning	None	0.63	0.57
Weakly-supervised	10%	0.62	0.56
	100%	**0.73**	**0.58**
Self-supervised	None	0.62	0.57

Table 5.2: ROI classification performances for different patch-level representation learning methods followed by a GNN-based aggregation.

Method	Label fraction	Accuracy	Bal. Accuracy
Transfer Learning	None	0.65	0.51
Weakly-supervised	10%	0.53	0.48
	100%	**0.74**	**0.61**
Self-supervised	None	**0.69**	**0.55**

accuracy increases with the number of labeled examples supports this analysis. Hence, we do not recommend combining self-supervised learning in the patch-level with a simplistic ROI-level aggregation method.

On the other hand, the GNN on top of patch features further models the diagnostically-relevant tissue features. In that case, the self-supervised features which do not carry the signal from the ROI labels outperform transferred features and weakly-supervised features in the low-label setting. We suggest that the choice between the weakly-supervised and self-supervised methods should depend on the ratio of the number of labeled examples to the number of unlabeled examples in a given dataset. When the data labeling cost is orders of magnitude higher than the data collection cost, utilizing the unlabeled data through self-supervised learning is beneficial.

5.1.3 Graph Neural Networks

We investigate the benefit of modeling ROIs with graph neural networks (GNNs) by comparing four GNN models with average pooling followed by an MLP, and a

state-of-the-art attention-based multiple instance learning method for ROI classification task. We also investigate the effects of different graph convolution and pooling operators by comparing four GNN variants. For this comparison, we use the same patch features extracted using the proposed self-supervised learning method, and train the ROI-level models in a supervised fashion using the cross-entropy loss function.

5.1.3.1 Model Configurations and Training

The method described in Section 5.1.2.1 is used to extract patches and build ROI graphs. A self-supervised trained patch-level model is used to compute the representations of the extracted patches.

The first method (Avg) computes the average of feature representations of all patches which comprise a region of interest. Then, a two-layer MLP classification head with a hidden dimension of 512 is used to retrieve the class scores for an ROI. Parameters of the MLP are initialized randomly using Kaiming-uniform method. The model is optimized using cross-entropy loss with a batch size of 32. Adam optimization algorithm with corrected weight decay regularization is used with 1×10^{-6} weight decay factor. One-cycle learning rate scheduling policy is employed with 20000 iterations warm up and 1×10^{-3} maximum learning rate.

The second, attention-based multiple instance learning (Attention) method is adopted from [68] which was described in Section 4.3.1. The hidden dimension of the attention layer, L in Equation 4.4, is set to 128. The same weight initialization, optimization, and scheduling methods with the previous model are used.

The first GNN variant (GNN-1) consists of two layers of graph convolutions defined in Equation 4.6 without local pooling. The hidden dimensions of the graph convolution layers are 128. The output features are globally averaged over all vertices to produce the ROI-level representation. Then, a 2-layer MLP with a hidden dimension of 64 is used to produce 4 logits for four-class classification.

Table 5.3: ROI classification performances for different ROI-level aggregation models with a fixed patch-level feature extractor.

Method	Accuracy	Bal. Accuracy
Avg	0.62	0.56
Attention	0.64	0.55
GNN-1	**0.69**	**0.58**
GNN-1-P	0.67	0.56
GNN-2	0.68	0.56
GNN-2-P	0.63	0.57

The second GNN variant (GNN-1-P) contains the same convolution operator with GNN-1, top-k pooling operators defined in Section 4.3.2 are inserted between the convolution layers. The best performance was achieved using 3 layers of convolutions and 2 pooling operations.

The third GNN variant (GNN-2) uses the second type of graph convolution defined in Equation 4.7, and no pooling. Two layers of convolutions with hidden dimensions of 128 provided the best result.

The forth GNN variant (GNN-2-P) uses the same graph convolution with GNN-2, but adds top-k pooling after each convolution. The maximum and average of vertex features after each pooling are concatenated, and passed through a two-layer MLP as described in [40] to obtain the ROI representation.

All of the four GNN models are trained using cross-entropy loss with a batch size of 16.

5.1.3.2 Results

The comparison of the ROI classification performances of the methods is given in Table 5.3. The results show that GNN-1 outperforms other methods by generating more efficient ROI representations. They also show that the GNN architecture and the choice of the graph convolution operator significantly effects the classification performance. We argue that the top-k pooling operator which

enables learning hierarchical features did not work well in our experiments because supervised training with a small dataset favored shallow GNN models with two convolution layers to the deep GNN models. Further experiments with other datasets are required for a better comparison. We suggest using shallow GNNs at the ROI level while using supervised learning with a small histopathology dataset, and experimenting with different graph convolution operators.

5.1.4 Self-Supervised Learning

In this set of experiments, the benefits of using self-supervised pre-training are explored. We use 100% of the examples in the dataset for self-supervised pre-training, and 10% of the examples for supervised fine-tuning. We compare this with supervised-only learning with 100% of the examples and with 10% of the examples. We also compare the proposed SimCLR-based method with another state-of-the-art self-supervised learning method called Momentum Contrast (MoCo) [49, 50].

5.1.4.1 Model Configurations and Training

A subset of the training set is created and fixed by randomly choosing 10% of the examples. The GNN models described in Section 5.1.3 are used with the same configurations. The labeled training set and 10% subset are used for supervised training of the models. Separately, the models are pre-trained with the method proposed in Section 4.3.3 using the original training set but without labels. For this step, the MLP classification heads are replaced with projection heads with a hidden dimension of 128 and output dimension of 32. The encoder and the projection head are optimized end-to-end using NT-Xent loss defined in Equation (4.1) with a batch size of 32 which results in 64 ROIs after the augmentation step. The loss temperature hyperparameter is set to 0.1. After pre-training is completed, the projection heads are removed and classification heads are added. Then, the classification heads are trained and the GNNs are fine-tuned using the

Table 5.4: ROI classifications results for different training schemes.

Method	Label fraction	Architecture	Accuracy	Bal. Accuracy
Supervised	10%	GNN-1	0.44	0.36
		GNN-1-P	0.50	0.47
		GNN-2	0.50	0.47
		GNN-2-P	0.49	0.48
Supervised	100%	GNN-1	**0.69**	**0.58**
		GNN-1-P	0.67	0.56
		GNN-2	0.68	0.56
		GNN-2-P	0.63	0.57
SSL-MoCoV2	10%	GNN-1	0.57	0.44
		GNN-1-P	0.56	0.43
		GNN-2	0.56	0.44
		GNN-2-P	0.55	0.44
SSL-SimCLR	10%	GNN-1	0.63	0.53
		GNN-1-P	0.64	**0.54**
		GNN-2	0.64	**0.54**
		GNN-2-P	**0.65**	0.53

labeled training subset.

The experiments are repeated by employing MoCo in the pre-training step. The encoder momentum is set to 0.999, temperature to 0.07, and batch size to 16.

5.1.4.2 Results

The results in Table 5.4 show that the proposed method achieves a classification performance similar to supervised learning with 674 (100%) labeled examples, but without needing any labels. When we reduce the number of labeled examples to 67 (10%) the classification accuracy drops by 19% on average. With the supervised method, the smaller GNN variants performed better than the larger ones; but the gap between them closes when self-supervised method is used. We infer larger models would benefit more from self-supervised pre-training as other studies on self-supervised learning also observed. For all GNN variants, SimCLR method consistently performed better than the MoCo.

Table 5.5: Per-class ROI classification performances of supervised learning with 10% of the labels (SL-10), with 100% of the labels (SL-100), and self-supervised pre-training supervised fine-tuning with 10% of the labels (SSL).

Method	Class	TP	FP	FN	TN	Precision	Recall	F_1
SL-10	Benign	120	99	27	126	0.55	0.82	0.65
	Atypia	10	20	28	314	0.33	0.26	0.29
	In Situ	9	18	119	226	0.33	0.07	0.11
	Invasive	44	52	15	261	0.46	0.74	0.57
SL-100	Benign	122	74	25	151	0.62	0.83	0.71
	Atypia	11	25	27	309	0.31	0.29	0.30
	In Situ	65	26	63	218	0.71	0.51	0.59
	Invasive	34	15	25	298	0.69	0.58	0.63
SSL	Benign	130	82	17	143	0.61	0.88	0.72
	Atypia	10	17	28	317	0.37	0.26	0.31
	In Situ	70	24	58	220	0.75	0.55	0.63
	Invasive	30	9	29	304	0.77	0.51	0.61

The per-class classification metrics of the GNN-2-P model are given in Table 5.5. SL-10 and SL-100 indicate supervised learning with 10% and 100% of the labeled examples, respectively. SSL indicates the proposed SimCLR-based self-supervised learning method followed by supervised fine tuning with 10% of the labeled examples. The results show that SL-100 and SSL models behave similarly, but SSL outperforms SL-10 in every class. Among the classes, Atypia has the lowest classification accuracy in all models, which we attribute to its small example count in the dataset. The performance difference between SL-10 and SSL is the most significant in the In Situ class. It is generally more difficult to discriminate the In Situ class than the two extreme classes which are Benign and Invasive. We infer that the self-supervised representation learning method can be more beneficial in fine-grained multi-class histopatology image classification setting. Overall, the results support our hypothesis that the self-supervised representation learning on graph models of ROIs is capable of generating efficient ROI representations.

Table 5.6: ROI classification performances for different patch sizes.

Patch Size	Accuracy	Bal. Accuracy
64	**0.63**	**0.53**
128	0.62	0.52
256	0.50	0.39

5.1.5 Effects of the Patch Size

The patch size is an important hyperparameter for setting the context for the model while separating the learning process into two stages. To investigate its effect on the classification performance, we train three patch-level and three ROI-level models using patches with 64, 128, and 256 pixels size at each dimension. The overlap between the patches during the extraction process is kept at 50%, that is the step size of the sliding window is 32, 64, and 128, respectively. Using patches smaller than 64 pixels size renders ROI-level model training infeasible with our hardware resources, while using patches larger than 256 pixels discards too many ROIs with less than 256 pixels width or height. We apply the proposed two stage method with self-supervised patch-level learning and self-supervised ROI-level pre-training with supervised fine tuning. We employ the GNN-1 architecture which was described in Section 5.1.3.

The results are given in Table 5.6. We observe that the classification performance decreases with the increasing patch size. Considering the content of a patch at 10× magnification level with the changing patch size, the result supports the fundamental idea behind the proposed two stage method in which the patch-level model focuses on cell-level features and tissue-level structures are modeled with the GNN. Further performance improvements are possible when working with smaller patches and/or higher magnification levels if sufficient compute and memory resources are available.

5.2 Retrieval

In addition to the classifier systems providing diagnostic predictions or grading scores, content-based image retrieval (CBIR) has also been investigated for decision support in many clinical applications [80, 81, 82, 83]. Given an image database, CBIR methods aim to retrieve images with morphological characteristics most relevant to and consistent with the query image. CBIR can also be used for classification purposes by considering the majority diagnosis of the retrieved images as the most likely diagnosis. State-of-the-art results in CBIR are achieved using supervised representation learning through classification. However, their power is bounded by the amount of manually labeled training data. Transfer learning and autoencoder-based unsupervised learning are also explored, but their success in large-scale images remain limited [26, 27, 28].

We propose using self-supervised representation learning method defined in Chapter 4 for content-based histopathology image retrieval task. The proposed method is compared with a state-of-the-art supervised learning method [19]. GNN-based architectures listed in Section 4.3.2 are used in the experiments.

5.2.1 Model Configurations and Training

We utilize GNN-1 and GNN-2-P architectures which were described in Section 5.1.3.1. In addition, we employ the DiffPool architecture [41] as GNN-3. DiffPool learns differentiable dense cluster assignments for vertices at each layer, mapping each vertex to a cluster. Each cluster becomes the input vertex for the next layer. The GNN-3 variant has two pooling operations that divide the network into three hierarchical layers. The first two layers have one subnetwork for vertex features and one subnetwork for cluster assignments. The last layer has only one subnetwork for vertex features. Each subnetwork consists of two consecutive SAGE convolution layers [84].

For supervised learning, a two-layer MLP with a hidden dimension of 128 and

output dimension of 4 is appended to the GNN encoder. For a fair comparison, the vertex dropout augmentation is applied to the training set in supervised learning. The GNN and MLP are trained end-to-end using cross-entropy loss function.

We compare our methods with GNN-LR [19], which is a state-of-the-art method for classification-based representation learning. The only modification we made is to use overlapping instead of non-overlapping sliding window approach used in the paper. The dense sampling method performed better than the non-overlapping alternative in both supervised and self-supervised settings.

All models are trained using a variant of Adam optimizer with weight decay proposed in [77]. The parameters of the optimizer, the temperature parameter in the loss function, the hidden layer sizes, and pooling ratios are determined through hyperparameter optimization based on the validation set performance. The experiments are implemented in Python language using PyTorch machine learning framework and Lightning research framework. All training tasks are carried out using a single machine with 2 NVIDIA RTX 2080Ti GPUs.

5.2.2 Evaluation Method

The outputs of the GNN encoders in both methods are used as the ROI representations for content-based image retrieval. In all experiments, the same training, validation, and test sets are used for model training, model selection, and retrieval performance estimation, respectively. The training and validation sets are included neither in the gallery nor in the query sets in the CBIR setup for a realistic evaluation. The test set is randomly split into query and gallery subsets according to 1-to-4 ratio and the same split is used in all experiments.

For each ROI in the query set, regions in the gallery set are retrieved in a ranked order based on the Euclidean distance between their representations. Regions that have the same class label with the query are considered relevant results while others are considered irrelevant. Mean Average Precision (mAP) metric is

used for quantitative evaluation of retrieval performance. Average Precision (AP) is the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight as

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{5.1}$$

where P_n is the precision and R_n is the recall at the n'th threshold. Then, the mAP is defined as the mean of average precisions over all queries as

$$mAP = \frac{1}{Q} \sum_{q=1}^{Q} AP(q). \tag{5.2}$$

AP@K is defined as the AP calculated using the top K retrieved items. MAP@K is calculated using AP@K. Considering the size of the test set, MAP@10 and MAP@25 are calculated and reported in the experiments.

5.2.3 Results

The quantitative results are summarized in Table 5.7. Example retrieval results are presented in Figure 5.1. The proposed self-supervised method outperforms GNN-LR in all experiments. Unlike the classification task, there is no supervised fine tuning step in the proposed retrieval method. Therefore, a content-based histopathology image retrieval system which has better performance than the state-of-the-art can be established without the labeling process. The comparison of GNN variants suggests simpler architectures may be preferred in this task.

Table 5.7: ROI retrieval performances for different architectures and training methods.

Method	Supervision	Architecture	MAP@10	MAP@25
GNN-LR [19]	Supervised	GNN-3	0.62	0.59
		GNN-2-P	0.73	0.64
		GNN-1	0.80	0.76
Ours	Self-Supervised	GNN-3	0.78	0.70
		GNN-2-P	0.82	0.75
		GNN-1	**0.86**	**0.80**

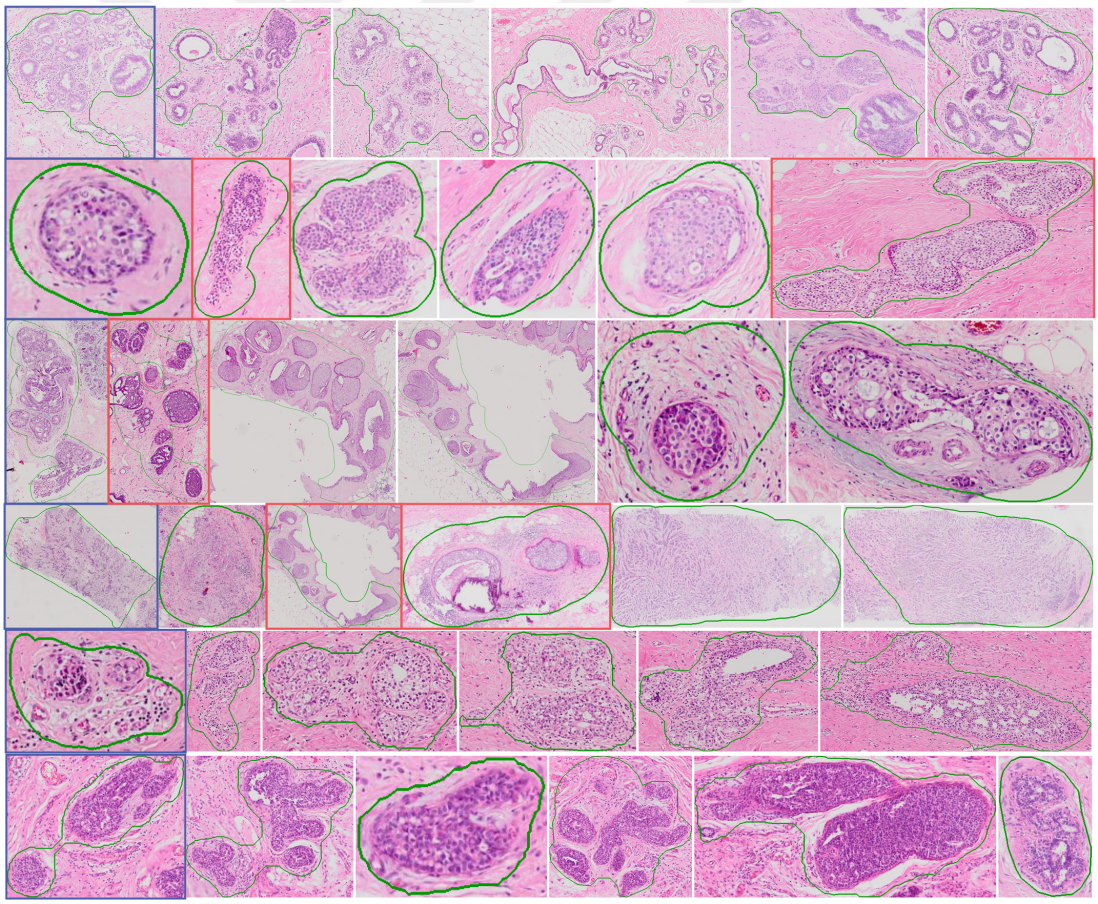

Figure 5.1: Examples of ROI retrieval using the best model, showing the top 5 retrieved items for each query in separate rows. Query images are marked in blue. Among the retrieved images, the irrelevant ones are marked in red. The ROI boundaries are marked in green. Query ROI classes from top to bottom are: benign, in-situ, in-situ, invasive, atypia, and benign.

Chapter 6

Conclusions and Future Work

In this thesis, we proposed a novel histopathology representation learning method. A two-stage learning pipeline is applied for separate patch-level and ROI-level representation learning. The regions with variable sizes an shapes are represented as graph, and graph neural networks are employed to encode contextual information. Contrastive self-supervised learning is used to combat scarcity of annotated samples.

Overall, we show that the proposed self-supervised learning method can learn effective representations in both patch and region of interest level despite not utilizing class labels. This result is consistent with our view that the challenge of histopathology dataset annotation can be tackled by contrastive self-supervised learning.

In separate sets of experiments, we showed that modeling ROIs as graphs improves the quality of learned representations. The self-supervised learning method demonstrated its advantages in both patch-level and ROI-level learning steps when the labeling operation is laborious and labeled datasets are scarce. For classification, the self-supervised model which is fine-tuned with 10% of the ROI labels has comparable performance to the supervised model which is trained with 100% of the labels. The GNN models in ROI-level are kept shallow due

to GPU memory constrains. We believe the self-supervised method would gain more if larger GNN models were trained on the same dataset. Self-supervised pre-training of the GNN models increased the classification accuracy by up to 20% when the labeled data for supervised learning is limited.

The proposed method for content-based histopathology image retrieval, without using class labels, has achieved better retrieval performance in a realistic breast histopathology dataset than its alternatives that use the same amount of data with class labels. Thus, the method allows utilizing the vast amount of unlabeled histopathology image data. The method enables cheaply training histopathology retrieval systems that can process arbitrarily-shaped queries.

In the future work, combining patch-level features from multiple magnification levels could be studied. Investigating the effects of using different graph augmentations in the proposed self-supervised method would also be worthwhile. Finally, by increasing the hardware resources, or utilizing more efficient model training techniques, the two stages in the pipeline can be merged, and the connected network can be trained in an end-to-end fashion.

Bibliography

[1] Z. Li, X. Zhang, H. Müller, and S. Zhang, "Large-scale retrieval for medical image analytics: A comprehensive review," *Medical Image Analysis*, vol. 43, pp. 66–84, Jan. 2018.

[2] M. Y. Lu, R. J. Chen, J. Wang, D. Dillon, and F. Mahmood, "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding," *arXiv preprint arXiv:1910.10825*, 2019.

[3] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1476–1485, 2019.

[4] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

[5] D. Bacciu, F. Errica, A. Micheli, and M. Podda, "A Gentle Introduction to Deep Learning for Graphs," *arXiv:1912.12693*, Dec. 2019.

[6] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[7] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*, vol. 9041, p. 904103, International Society for Optics and Photonics, 2014.

[8] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2560–2567, IEEE, 2016.

[9] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2440–2445, IEEE, 2016.

[10] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, N. Karssemeijer, G. Litjens, and J. van der Laak, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *Journal of Medical Imaging*, vol. 4, no. 4, p. 044504, 2017.

[11] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-net: joint segmentation and classification for diagnosis of breast biopsy images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 893–901, Springer, 2018.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[13] K. Das, S. P. K. Karri, A. G. Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 1024–1027, IEEE, 2017.

[14] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, 2016.

[15] P. Kainz, M. Pfeiffer, and M. Urschler, "Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization," *PeerJ*, vol. 5, p. e3874, 2017.

[16] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[17] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2424–2433, 2016.

[18] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "From patch-level to roi-level deep feature representations for breast histopathology classification," in *Medical Imaging 2019: Digital Pathology*, vol. 10956, p. 109560H, International Society for Optics and Photonics, 2019.

[19] Y. Zheng, B. Jiang, J. Shi, H. Zhang, and F. Xie, "Encoding histopathological wsis using gnn for scalable diagnostically relevant regions retrieval," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 550–558, Springer, 2019.

[20] J. Wang, R. J. Chen, M. Y. Lu, A. Baras, and F. Mahmood, "Weakly supervised prostate tma classification via graph convolutional networks," *arXiv preprint arXiv:1910.13328*, 2019.

[21] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 642–653, 2009.

[22] Y. Zheng, Z. Jiang, F. Xie, H. Zhang, Y. Ma, H. Shi, and Y. Zhao, "Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification," *Pattern Recognition*, vol. 71, pp. 14 – 25, 2017.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009.

[24] S. Vesal, N. Ravikumar, A. Davari, S. Ellmann, and A. Maier, "Classification of breast cancer histology images using transfer learning," in *International Conference Image Analysis and Recognition*, pp. 812–819, Springer, 2018.

[25] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1–6, 2019.

[26] T. Schlegl, J. Ofner, and G. Langs, "Unsupervised pre-training across image domains improves lung tissue classification," in *International MICCAI Workshop on Medical Computer Vision*, pp. 82–93, Springer, 2014.

[27] B. Kieffer, M. Babaie, S. Kalra, and H. 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)*, pp. 1–6, IEEE, 2017.

[28] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran, *et al.*, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images," *Pattern Recognition*, vol. 86, pp. 188–200, 2019.

[29] J. A. Vanegas, J. Arevalo, and F. A. Gonzalez, "Unsupervised feature learning for content-based histopathology image retrieval," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, (Klagenfurt, Austria), pp. 1–6, IEEE, June 2014.

[30] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 496–506, Feb. 2015.

[31] Y. Ma, Z. Jiang, H. Zhang, F. Xie, Y. Zheng, H. Shi, Y. Zhao, and J. Shi, "Generating region proposals for histopathological whole slide image retrieval," *Computer Methods and Programs in Biomedicine*, vol. 159, pp. 1–10, June 2018.

[32] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Histopathological Whole Slide Image Analysis Using Context-Based CBIR," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1641–1652, July 2018.

[33] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Size-Scalable Content-Based Histopathological Image Retrieval From Database That Consists of WSIs," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 1278–1287, July 2018.

[34] Y. Zheng, B. Jiang, J. Shi, H. Zhang, and F. Xie, "Encoding Histopathological WSIs Using GNN for Scalable Diagnostically Relevant Regions Retrieval," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), vol. 11764, pp. 550–558, Cham: Springer International Publishing, 2019.

[35] A. Sperduti and A. Starita, "Supervised neural networks for the classification of structures," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 714–735, 1997.

[36] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2013.

[37] J. Atwood and D. Towsley, "Diffusion-convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1993–2001, 2016.

[38] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International Conference on Machine Learning*, pp. 2014–2023, 2016.

[39] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv:1609.02907*, Feb. 2017. arXiv: 1609.02907.

[40] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and leman go neural: Higher-order graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4602–4609, 2019.

[41] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical Graph Representation Learning with Differentiable Pooling," *arXiv:1806.08804*, Feb. 2019. arXiv: 1806.08804.

[42] J. Wang, R. J. Chen, M. Y. Lu, A. Baras, and F. Mahmood, "Weakly supervised prostate tma classification via graph convolutional networks," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 239–243, IEEE, 2020.

[43] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, "Graph neural network for interpreting task-fmri biomarkers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 485–493, Springer, 2019.

[44] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[45] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*, pp. 649–666, Springer, 2016.

[46] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*, pp. 69–84, Springer, 2016.

[47] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. v. d. Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding," *arXiv:1905.09272*, Dec. 2019.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv:2002.05709*, Feb. 2020. arXiv: 2002.05709.

[49] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[50] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[51] J. Gildenblat and E. Klaiman, "Self-supervised similarity learning for digital pathology," *arXiv preprint arXiv:1905.08139*, 2019.

[52] A. C. Ruifrok, D. A. Johnston, *et al.*, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[53] A. Basavanhally and A. Madabhushi, "Em-based segmentation-driven color standardization of digitized histopathology," in *Medical Imaging 2013: Digital Pathology*, vol. 8676, p. 86760G, International Society for Optics and Photonics, 2013.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[55] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[56] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 793–802, IEEE, 2018.

[57] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[58] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, p. 101539, 2019.

[59] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[61] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[62] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," *arXiv preprint arXiv:1912.11370*, vol. 6, p. 3, 2019.

[63] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[64] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[65] C. Mercan, *Deep feature representations and multi-instance multi-label learning of whole slide breast histopathology images*. PhD thesis, Bilkent University, 2019.

[66] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[67] C. Mercan, E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for whole slide breast histopathology," in *Medical Imaging 2016: Digital Pathology*, vol. 9791, p. 979108, International Society for Optics and Photonics, 2016.

[68] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[69] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[70] H. Gao and S. Ji, "Graph u-nets," *arXiv preprint arXiv:1905.05178*, 2019.

[71] C. Cangea, P. Veličković, N. Jovanović, T. Kipf, and P. Liò, "Towards sparse hierarchical graph classifiers," 2018.

[72] B. Knyazev, G. W. Taylor, and M. Amer, "Understanding attention and generalization in graph neural networks," in *Advances in Neural Information Processing Systems*, pp. 4204–4214, 2019.

[73] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[74] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[75] W. Falcon, "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, 2019.

[76] L. Melas-Kyriazi, "Efficientnet-pytorch," *GitHub. Note: https://github.com/lukemelas/EfficientNet-PyTorch*, 2019.

[77] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[78] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, p. 1100612, International Society for Optics and Photonics, 2019.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings*

of the IEEE International Conference on Computer Vision, pp. 1026–1034, 2015.

[80] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[81] H. C. Akakin and M. N. Gurcan, "Content-based microscopic image retrieval system for multi-image queries," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 758–769, 2012.

[82] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 496–506, Feb. 2015.

[83] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Histopathological Whole Slide Image Analysis Using Context-Based CBIR," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1641–1652, July 2018.

[84] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *arXiv:1706.02216*, Sept. 2018. arXiv: 1706.02216.