MODELING SPATIAL CONTEXT IN TRANSFORMER-BASED WHOLE SLIDE IMAGE CLASSIFICATION

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
Cihan Erkan
September 2023

Modeling Spatial Context in Transformer-based Whole Slide Image Classification
By Cihan Erkan
September 2023

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

Ayşegül Dündar

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

MODELING SPATIAL CONTEXT IN TRANSFORMER-BASED WHOLE SLIDE IMAGE CLASSIFICATION

Cihan Erkan
M.S. in Computer Engineering
Advisor: Selim Aksoy
September 2023

The common method for histopathology image classification is to sample small patches from the large whole slide images and make predictions based on aggregations of patch representations. Transformer models provide a promising alternative with their ability to capture long-range dependencies of patches and their potential to detect representative regions, thanks to their novel self-attention strategy. However, as sequence-based architectures, transformers are unable to directly capture the two-dimensional nature of images. Modeling the spatial context of an image for a transformer requires two steps. In the first step the patches of the image are ordered as a 1-dimensional sequence, then the order information is injected to the model. However, commonly used spatial context modeling methods cannot accurately capture the distribution of the patches as they are designed to work on images with a fixed size. We propose novel spatial context modeling methods in an effort to make the model be aware of the spatial context of the patches as neighboring patches usually form diagnostically relevant structures. We achieve that by generating sequences that preserve the locality of the patches. We test the generated sequences by utilizing various information injection strategies. We evaluate the performance of the proposed transformer-based whole slide image classification framework on a lung dataset obtained from The Cancer Genome Atlas. Our experimental evaluations show that the proposed sequence generation method that utilizes space-filling curves to model the spatial context performs better than both baseline and state-of-the-art methods by achieving 87.6% accuracy.

Keywords: Digital pathology, space-filling curves, vision transformer, whole slide image classification.

ÖZET

DÖNÜŞTÜRÜCÜ TABANLI TÜM SLAYT SINIFLANDIRMASINDA UZAYSAL BAĞLAMIN MODELLENMESI

Cihan Erkan
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Selim Aksoy
Eylül 2023

Histopatolojik görüntü sınıflandırmasının en yaygın yöntemi, büyük tüm slayt görüntülerinden toplanan küçük pencerelerin gösterimlerini birleştirerek bir tahmin yapmaktır. Dönüştürücü modelleri, özgün öz-dikkat stratejileri sayesinde pencereler arasındaki uzak mesafe ilişkilerini anlayabildikleri ve önemli bölgeleri saptayabildikleri için bu alanda dikkat çeken bir alternatiftir. Ancak dizi tabanlı mimarileri yüzünden görüntülerin iki boyutlu doğasını doğrudan anlayamazlar. Bir tüm slaytın uzamsal bağlamı iki aşamada modellenir. İlk aşamada pencereler tek boyutlu bir dizi halinde sıralanır, daha sonra bu sıra bilgisi modele enjekte edilir. Ancak, yaygın kullanılan uzamsal bağlamı modelleme yöntemleri sabit boyutlardaki görüntüler için tasarladıkları için bu pencerelerin görüntü içindeki dağılımını tam anlamıyla anlayamazlar. Bu pencereler bir araya gelerek teşhis bakımından anlamlı yapılar oluşturabilecekleri için, pencerelerin uzamsal dağılımını anlayabilecek bir uzamsal bağlamı modelleme yöntemi öneriyoruz. Bunu yerelliği koruyan diziler oluşturarak başarıyoruz. Oluşturulan dizileri farklı bilgi enjekte etme yöntemleriyle test ettik. Önerdiğimiz dönüştürücü tabanlı tüm slayt sınıflandırma yönteminin başarımını The Cancer Genome Atlas'ın akciğer kanseri veri kümesinde değerlendirdik. Deneysel sonuçlarımıza göre, boşluk dolduran eğrileri kullanan yöntemimiz %87.6 doğruluğa erişerek, temel ve modern yöntemlerden daha yüksek başarım göstermiştir.

Anahtar sözcükler: Dijital patoloji, boşluk dolduran eğriler, görüntü dönüştürücü, tüm slayt görüntüsü sınıflandırma.

Acknowledgement

I would like to express my gratitude to my advisor Prof. Dr. Selim Aksoy for his patience and kindness. None of these would have been possible without his invaluable guidance. His passion for research really inspired me to achieve my goals.

I would also like to thank Assoc. Prof. Dr. R. Gökberk Cinbiş and Asst. Prof. Dr. Ayşegül Dündar for accepting my invitation to my defense and their valuable feedback on my thesis.

Words cannot express my gratitude to Banu for always being there whenever I needed her. Her endless love and support made every step of this journey much more pleasant.

This would have been much more difficult without the unconditional support of my family. They always believed in me and encouraged me to move forward.

Finally, I would like to thank the Scientific and Technological Research Council of Turkey (TUBITAK) for providing financial support for my master's studies with the grant 117E172.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diagnosing cancer requires a pathologist to examine the samples collected from a patient. This process is exceedingly important as it is the main factor that determines the method of treatment. However, it is also an exceptionally tedious task, requiring careful inspection of multiple biological structures. Fortunately, these samples can be digitized to obtain whole slide images (WSIs). This process paves the way for utilizing computer based tools to analyze WSIs, which in turn eases off the workload of the pathologists.

In their essence, WSIs are giga-pixel sized images. Consequently, it is possible to formulate the cancer diagnosis task as an image classification problem. Even though image classification is a thoroughly studied problem in computer science, WSI classification is not a straightforward task. That is because WSIs have some unique attributes that are not present in broadly studied image types.

The most eye-catching difference of WSIs are their tremendous size. WSIs can easily have resolutions larger than $100,000 \times 100,000$ pixels which makes it impossible to use commonly used image classification methods, such as convolutional neural networks (CNNs). Another distinct feature of the WSIs is their variable shapes. Tissues in WSIs can have highly elongated shapes. This poses a serious challenge, as common image classification frameworks usually require

images to have a fixed shape. Normally, this issue can be sidestepped by resizing or cropping the image. However, such simple solutions cannot handle the aforementioned extreme shape variations in tissue structures.

Earlier efforts to solve the WSI classification problem usually followed the footsteps of pathologists. Since pathologists diagnose the cancer by evaluating the biological structures present in the sample, these methods tried to extract biologically meaningful features from WSIs [1, 2, 3]. These features mainly capture the information related to cell morphology and their distribution. When used in conjunction with machine learning models like support vector machines (SVMs), they were capable of producing acceptable results in WSI classification tasks.

Most modern tools usually follow strategies that revolve around patches. Here, the WSI is divided into fixed sized patches. This approach negates the downside of WSIs, as their larger sizes and irregular shapes only affect the total number of patches extracted from a WSI. Moreover, since patch sizes are fixed, it is possible to use common pre-trained networks to extract deep features from these patches. The actual classification is done by aggregating these features. Different WSI classification methods usually diverge from each other in this step, based on the method they follow to combine patch features to obtain a WSI-wide representation.

Advancements in WSI classification methods cannot be separated from the innovations in image classification frameworks. Transformers, even though they were initially proposed as a natural language processing (NLP) model [4], proved to be a promising solution to image classification as well [5]. Being an NLP architecture, they work on sequences of tokens which is quite intuitive for natural language sentences, though this is not the case for 2-dimensional images. Nevertheless, it is possible to represent an image as a sequence of tokens by dividing it into smaller patches and ordering them as a 1-dimensional sequence. However, this step unveils two major weaknesses of transformer based architectures.

One of the main drawbacks of transformers is their quadratic computing cost, especially for the WSIs as their enormous sizes create considerably long patch

sequences. Due to that, it is almost inevitable to use more lightweight variants of the transformer architecture. These models either limit the global attention of transformers [6], or approximate self-attention [7]. Another important problem with the transformer architecture is their inability to understand the ordering of a sequence. While transformers can still process a sequence without knowing its order, it is vital to inject the position information to the model as it has been shown to boost the performance substantially [4]. The most commonly used strategy here is adding a positional encoding vector to the tokens. This allows transformers to recognize the spatial relationships between the members of the sequence. Even though certain encoding methods can faithfully represent a 1-dimensional sequence, it becomes a more convoluted process when we try to model the 2-dimensional nature of the WSIs.

There are multiple ways of creating a sequence out of patches of an image. The most simple and straightforward approach is following the raster scan order. This method presents a lossless way of transferring the 2-dimensional information to a single dimension, but only when the shape and size of the image is fixed. If we apply the same method to a WSI, the resulting ordering provides almost no information about the spatial context of the patches due to their variable shapes and sizes. When patches that constitute a single diagnostically relevant structure are scattered around the sequence with no apparent pattern that relates them to each other, the performance of the model can diminish significantly.

Our contributions in this thesis are as follows. First, we propose novel sequence generation methods that employ space-filling curves. The proposed approach improves the classification performance of our transformer-based WSI classifier and it surpasses the performance of the state-of-the-art WSI classification models. Secondly, we experiment with multiple other sequence generation methods, and evaluate their performance. Lastly, we propose 2 new methods to model the spatial context of a WSI, one relative position information based approach that employs CNNs and one absolute position information method that directly models 2-dimensional distribution of the WSI patches. An earlier version of this work was published in [8].

The rest of this thesis is organized as follows. In Chapter 2, we discuss the related work regarding WSI classification, transformers and space-filling curves. Chapter 3 details the proposed transformer based WSI classification model, sequence generation and spatial context modeling methods. We present our experiments and their results in Chapter 4. Finally, Chapter 5 summarizes the findings of this thesis.

Chapter 2

Related Work

As the CNN models became the norm in image analysis problems, they replaced the previous methods that relied on feature engineering. Since they operate on fixed sized images, they require a patch based strategy on the WSI classification task. The prevalent approach first extracts feature vectors and then combines those vectors. The method of feature aggregation is the main difference among these models. For example, Zheng et al. [9] extracts patches from the viewing path of a pathologist, then uses a recurrent neural network (RNN) architecture to classify WSI. The method proposed by Campanella et al. [10] employs multiple instance learning (MIL) to train a patch feature extractor and an RNN to make a WSI-wide decision. Since it is costly to use all patches, Xu et al. [11] propose employing an RNN module which guides the patch sampling strategy while predicting the current WSI label to avoid uninformative patches. Vu et al. [12] propose a MIL model, where they train a model to learn patch labels using a logistic regression function. Then the proposed framework predicts the WSI label based on the labels of the patches that constitute the WSI. Furthermore, Ilse et al. [13] propose adding an attention layer on top of their MIL based method. That way, they were able to obtain a heatmap regarding the importance of each instance. Lu et al. [14] propose an attention based MIL model for multi-class classification of WSIs. Their work utilizes multiple attention branches for each of the classes present in their dataset. They also encourage the model to learn class

specific features by introducing a clustering objective, where the model learns to differentiate the informative and non-informative instances for a given class.

Vaswani et al [4] proposed using transformers as a machine translation model, capable of translating text from English to German and French. However, they proved themselves to be useful in many other fields as well. For example, Devlin et al. [15] proposed using transformers as the backbone of a language representation model. The authors argue that with proper fine-tuning, the model is capable of performing many tasks, such as question answering and next sentence prediction.

Dosovitskiy et al. [5] introduced the transformer architecture to the field of computer vision by proposing a transformer based solution to the image classification problem. Their method converts an image into a sequence by following a patch based strategy, which became the basis for most computer vision applications that utilize transformers. Wang et al. [16] proposed a model called TRPose for human pose estimation, where they combined CNN and transformer based architectures. Transformers can also be utilized in image generation tasks as well as demonstrated by Huang et al. [17]. The authors propose a U-Net-like transformer model to perform the image inpainting task.

As transformers became more prevalent in other areas of computer vision, the number of medical image analysis tools that utilize the transformer architecture surged as well. Roy et al. [18] proposed a transformer model to detect symptoms of COVID-19 in lung ultrasound images that can also predict the severity of the disease. In order to perform segmentation on 3-dimensional medical data, Hatamizadeh et al. [19] proposed a sequence-to-sequence transformer. The WSI classification task can be performed with transformer-based methods as well. These models too, usually follow patch based strategy similar to the ones used by CNN based methods. However, most of the proposed approaches, refrain from modeling the spatial context of the patches due to the challenges of representing WSIs as a sequence. This indicates that the state-of-the-art WSI classification frameworks can be improved by using a positional encoding strategy that is capable of modeling the spatial context of a WSI.

For example, Shao et al. [20] proposed a transformer based WSI classification model where the positional information is transferred to the classifier via a 2-dimensional CNN, placed between two transformer layers. They utilized this CNN layer to incorporate the position information. However, as the 2-dimensional CNNs require the patches to be ordered in a 2-dimensional manner, the method proposed re-arranges the patches in a square form. Considering the structure of the tissue is most likely not a perfect square, this process causes some information to be lost. Even though this method somewhat preserves the information in the horizontal axis, it almost completely discards the vertical axis information. Consequently the model only achieves subpar performance.

Chen et al. [21] employs a transformer based MIL model for survival prediction which utilizes both genomic and histopathologic data. Normally the self-attention mechanism of the transformers calculates the attention between the elements of the same structure, e.g. words of a sentence or patches of a WSI. However, the proposed model works by calculating the attention between different sets of data, i.e. the WSI and genomic attributes of the patient. The authors preprocess the WSI with a patch based approach, and use the patch features for their transformer architecture. However, they do not employ any method that can capture the spatial distribution of the patches within the WSI.

The model proposed by Gul et al. [22] aims to use the transformer as a feature extraction model. To train the transformer-based feature extractor they follow a weakly supervised strategy, where they divide the WSI into 224×224 pixel sized patches and assume all resulting patches share the WSI label. Then they predict the patch label using a transformer. The input tokens of the proposed model are the linear projections of 16×16 tiles cropped from the patch. They model the spatial context of a single patch using positional encodings. However, modeling the spatial context of a single, fixed sized patch is a relatively straightforward process. Even though they do not completely discard the positional information, the proposed approach fails to introduce a novel method which can capture the spatial arrangement of a WSI.

Mehta el al. [23] proposes a strategy where they utilize multiple transformers

which relate the patches cropped at different scales. The authors argue that incorporating positional encodings does not improve the performance of the model as the patches extracted in different scales implicitly model the spatial context. However, the positional encodings they experimented with do not address the WSI specific challenges. Because of that, the performance of the model can be increased by using a method that successfully models the position information of a WSI.

As mentioned previously, the self-attention mechanism of the transformers are permutation invariant. Because of that, transformers cannot understand the spatial context of the input tokens. However, this issue is usually addressed by using a special type of vector called positional encodings. These encodings, either learnable or constant, can be added to the token representations as a method of modeling the spatial context. While the learnable encodings [5] usually perform better, they require the length of the token sequence to be limited. On the other hand, sine-cosine based constant positional encodings proposed by [4] can model arbitrarily long sequences. Another method of modeling spatial context is employing relative position representations. Introduced by Shaw et al. [24], this method focuses on the relative distance between the tokens. While this causes the model to be unaware of the exact location of a specific token, it can still recognize its spatial context. However, all of these methods require tokens to be ordered as a 1-dimensional sequence. Because of that, their performance heavily depends on the sequence generation strategy.

Due to their locality preserving quality and ability to create a mapping between N-dimensional spaces and 1-dimensional sequences, space-filling curves are widely used in a number of fields. For example, Asano et al. [25] propose using space-filling curves to map multidimensional data to a disk. Since the physical locations of data blocks significantly impacts data retrieval speed, they argue utilizing space-filling curves improves the performance. Similarly, Li et al. [26] argue that such an approach is also useful for storing large geo-spatial images. Moreover, Chen et al. [27] note that since nearest neighbor finding is mostly concerned with proximity, space filling curves can be applied as they can preserve the locality. Furthermore, Biswas [28] proposes using space-filling curves to compress an image

file by exploiting the fact that pixels in similar locations have similar values. The mapping property of space-filling curves is useful for converting multidimensional data to a single dimension. However, the converse is also possible. Anders [29] utilizes space-filling curves to map 1 dimensional genomic data to an image for visualization purposes. In this thesis, we use the space-filling curves to generate patch sequences that preserve the locality of the patches.

Chapter 3

Methodology

Even though the transformer architecture was proposed as an NLP model, with slight modifications it can be utilized in different fields of computer science such as computer vision. Thus, the WSI classification problem, which is basically a special case of the image classification task, can be handled by transformer based methods. Moreover, their strong representation capacity and ability to understand long range dependencies gives them an edge over previously proposed, relatively simple models.

In the current literature, the patch based strategies are the norm for WSI classification. These methods define the WSI as a collection of patches, similar to how a sentence can be defined as a collection of words. This parallelism makes transformers an especially good method for WSI classification. With this approach, the task at hand becomes a sequence classification problem, where the elements of the sequence are the patches of a WSI. Consequently, the sequence based nature of transformers provides a natural way of transferring WSI patches to the transformer space. However, it is crucial to preserve the 2-dimensional ordering of the WSI during this transfer. This thesis proposes a transformer based WSI classification framework that is capable of understanding the ordering of WSI patches. Rest of this chapter details the transformer based architecture and the proposed method of modeling the spatial context.

3.1 Transformer Based Classification Model

In its essence, the transformer architecture is an encoder-decoder pair, proposed as a solution to the machine translation problem. Transformers operate on a sequence of instances, or tokens, and they output a different sequence of tokens. Even though it is a sequence-to-sequence model, when the transformer encoder is used on its own it can be the backbone of a sequence classification method.

Transformer encoders can create powerful representations of sequences. This is enabled by the multi-head self-attention (MHSA) layer as it is capable of generating a new set of tokens by relating each token to all other tokens. The self-attention mechanism can be mathematically formulated as:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

where Q, K, and V are a set of vectors generated from the input tokens and d_k is the size of Q and K. Q, K, and V vectors, commonly referred to as key, query and value, are learnable linear projections of the input token. Instead of learning a single linear projection, transformer models learn multiple projections and perform self-attention on different sets of Q, K, and V vectors to achieve "multi-headed" self-attention.

Because of the MHSA mechanism, transformers are a good fit for WSI classification. First of all, MHSA allows the model to relate each patch to every other patch. Because of that, a transformer based classifier can understand the long range relationships between patches. This is especially important for WSI classification as regions of interest can be scattered around the whole WSI. Furthermore, the attention based mechanism of this framework makes it possible to discard uninformative regions of the image. This provides a significant advantage to the model as it allows it to focus on the most important and meaningful parts of the sample. Another powerful tool of the transformer is its "multi-headed" approach which allows the model to understand different aspects of the data. With an input that is as complicated as the WSI, this tool becomes a significant advantage.

As powerful as they are, transformers are not free of glaring weaknesses. One of the most important drawbacks of the transformer architecture is its inability to understand the actual order of its tokens; or in the case of WSIs, its patches. This causes the transformer to be unaware of the greater context the patch resides in. As a result, transformers are blind to the structures that span multiple patches. Considering that these structures might be diagnostically relevant, it is essential to inject the spatial context of each patch to the model. For this reason, most of the modern transformer architectures utilize some form of position encoding to communicate the positions of tokens.

Providing the position information to the model can be done in many different ways. However, most of the time, this process requires two steps. The first step is generating a meaningful sequence to order the tokens. When the input is inherently single-dimensional, like sound or sentence, this step is trivial as there is a straightforward way of ordering the sequence. However, when the input is multi-dimensional, as is the case for images, there can be different methods of generating a sequence. After finding a sequence that is suitable for the data, the next step is modeling the spatial context of the input based on the sequence generated. Here the aim is to communicate the desired order to the transformer model. However, since the MHSA is permutation invariant, the order information cannot be directly inserted to the transformer. Instead, most architectures utilize an indirect method. The most common approach is slightly altering the input tokens based on their location in the sequence. It is important to note that the tokens actually carry valuable information regarding the input, and altering their values can cause them to be meaningless. However, when applied correctly, the aforementioned method can model the spatial context while preserving the meaning of the tokens.

Now with a strategy to introduce the position information to the model, it is possible to build a transformer based classification framework that can handle complex tasks like WSI classification. The method we propose in this thesis follows a patch based approach and can be decomposed into three components:

1. Feature extraction

2. Sequence generation and spatial context modeling

3. Classification

An overview of the proposed model is given in Figure 3.1.

The aim of the feature extraction step is generating a patch based representation of the WSIs. This allows us to create a more manageable representation of the original image. The first step of feature extraction is extracting patches from the WSI. We crop 256×256 non-overlapping patches. However, a significant portion of those patches are background patches and thus they provide no useful information regarding the label of the WSI. Continuing to process these patches would incur an unnecessary computation cost, so we discard such patches by employing a simple thresholding method. Usually the background of a WSI is a white-ish color, so we eliminate the patches whose average saturation is below the threshold. Then we extract the features of the remaining patches, using a ResNet model pre-trained on the ImageNet dataset. At the end of this step, we obtain a feature vector for each foreground patch of the WSI. In order to provide the spatial context, we order the foreground patches to generate a 1-dimensional sequence. Then we model the spatial context of the WSI using either absolute or relative position information schemes. Details of these steps are provided in Sections 3.2 and 3.3, respectively.

After extracting features and providing the spatial context, the final step is utilizing the tokens to make a prediction regarding the label of the WSI. However, since the transformer is a sequence-to-sequence architecture, it is not suitable for the classification task in its original form. Nevertheless, it is possible to work around this problem by discarding the decoder part and using a special token. This method converts the transformer to a sequence classification model. In this approach, the special token, called the class token, is a trainable token and it is inserted at the beginning of every sequence. After the sequence is processed by the transformer, the token that corresponds to the class token can be used for classification using an MLP head. We follow this method to predict the label of the WSI.

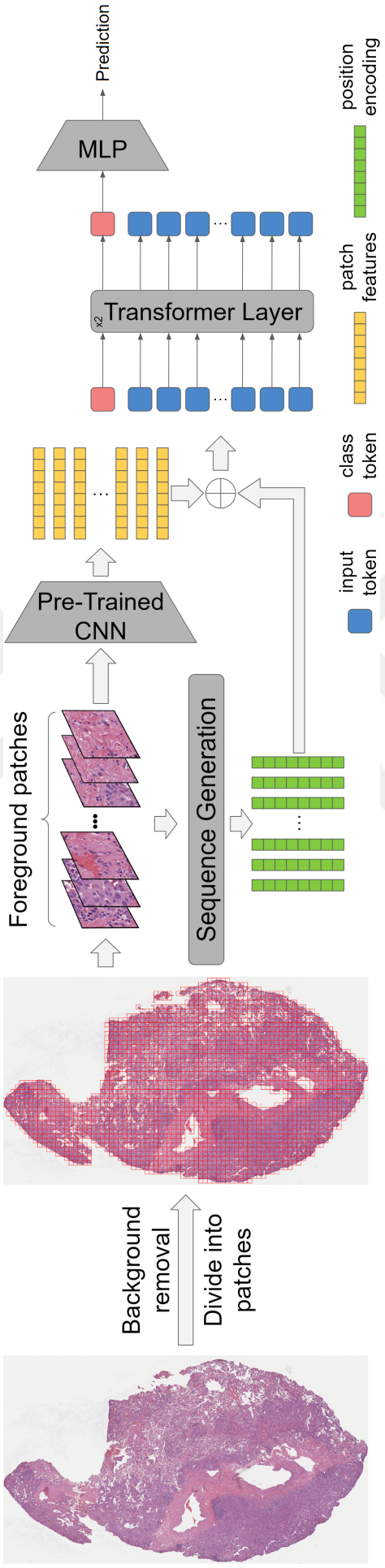


Figure 3.1: The overview of the proposed model.

Transformers require a significant amount of memory when the sequence length is large due to its quadratic space complexity. This poses a significant problem when they are used as a WSI classifier. Some patch sequences can include up to 60,000 tokens and consequently they do not fit into the memory of the GPU. We overcome this issue by using the Nyström method [7] which provides an approximation of the self-attention mechanism. This allows us to process all patches of the WSI at once.

The classification framework we propose includes a fully connected network that reduces the size of feature vectors to 512, two transformer layers with layer normalization in between and finally an MLP head attached to the output of the class token which predicts the label of the input WSI.

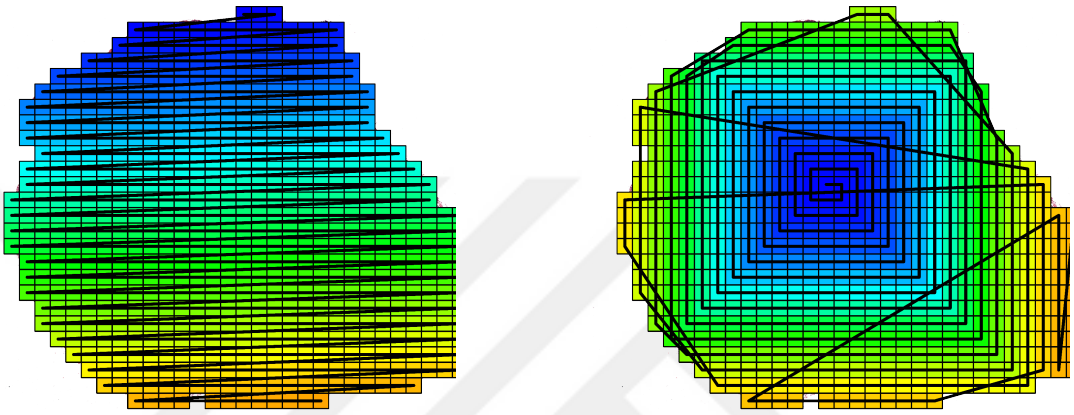
With the mentioned strategy, it is possible to classify WSIs using a transformer based classification network. However, as mentioned, supplying the position information to the model can be done in different ways. Rest of this chapter discusses different methods that can be utilized in this two-step process: Sequence generation and spatial context modeling, respectively.

3.2 Sequence Generation Methods

Since transformers are designed as a sequence based model, supplying the positional information of the patches involves generating a 1-dimensional sequence out of these patches. While it is possible to perfectly define a 2-dimensional image with a single dimensional ordering when the shape and size of the image is fixed, this is not the case for WSIs. Thus, it is essential to find a mapping between those two spaces that preserve as much information as possible.

Patches of a WSI can be ordered in a myriad of ways to obtain a single dimensional sequence. However, generating a sequence that also preserves the meaningful information regarding the actual distribution of the patches in the two-dimensional space requires following a certain methodology. For the purposes

of this thesis, we divided such sequence generation methods into three groups, namely, simple methods, space-filling curve based methods and data aware methods. Rest of this section discusses these methods in detail.



(a) (b)

Figure 3.2: The path traversed by the simple methods. (a) Scan, (b) Spiral. Each patch is drawn with a different color (by using a colormap from blue to orange) where neighboring patches have similar colors. The black lines also connect the patches in the order they appear in the sequence.

3.2.1 Simple Methods

Generating a sequence that captures the position information of the WSI does not necessarily require complex calculations. Certain simple methods can create orderings that reveal the underlying distribution of the patches relatively well. In this thesis we explore two such sequence generation strategies: Raster scan ordering and spiral ordering. Figure 3.2 illustrates the sequence generated by these methods.

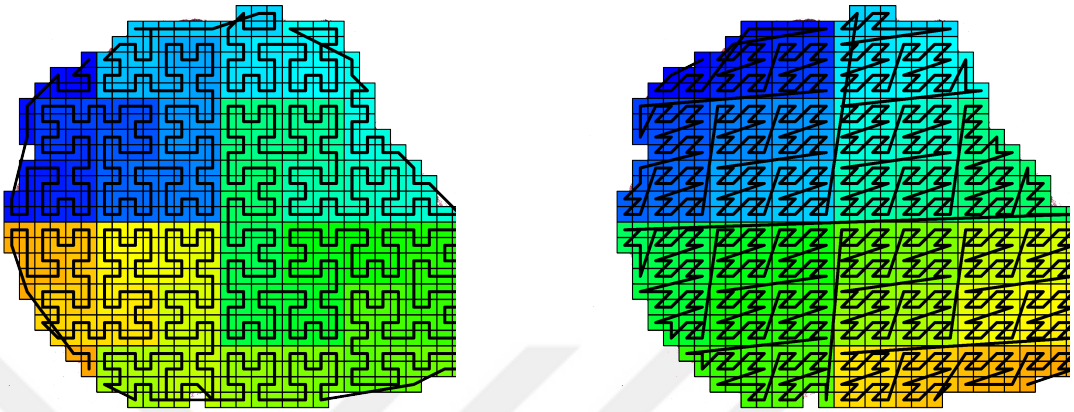
Raster Scan Ordering: The raster scan order is the most commonly used ordering method for the transformer based computer vision applications. The path it defines starts from the leftmost patch of the top row, and it travels left to right and top to bottom. Even though it is extremely simple, it preserves the location information, especially in the horizontal axis quite well. Moreover, when

the shape and size of the image is fixed, the exact position of each patch can be perfectly recovered from their location on the sequence generated by the raster scan order. However, applying this method to WSIs exposes its weaknesses. First of all, WSIs are not fixed sized images. Consequently, this method completely discards the vertical axis information as it is impossible to determine the neighbors of a patch that resides in a different row. Moreover, the background removal phase of the feature extraction step can introduce holes in the WSI, possibly breaking the continuity of the rows of patches. As a result, neighboring patches in the sequence space can be in wildly different regions in the image space, even when they occupy the same row. To sum up, while the raster scan order is commonly used, it may not be the best method when dealing with WSIs.

Spiral Ordering: The spiral ordering method aims to create a sequence that traverses a path defined by an outward traveling spiral that emanates from the center of the foreground region. With this method, the position information of the patches that occupy the central region of the WSI is preserved fairly well. Moreover, even though it is impossible to determine the exact location of a patch, at least its distance to the center region can be more or less predicted based on its position in the sequence. Lastly, considering the later sections of the spiral act similar to the scan order, the spiral ordering also embodies the inherent advantages of the raster scan order. However, this also means that the drawbacks of the scan order affects spiral order as well. Moreover, discontinuities in the foreground region, especially when located near the center, can drastically reduce meaningfulness of the sequence generated. Therefore, while this method offers some advantages over the scan order, it may not consistently perform better.

3.2.2 Space Filling Curves

Mathematically a space-filling curve can be defined as a line that passes through all points in a given space. Because of that, such a curve provides a one-to-one mapping between 1-dimensional and multidimensional spaces. Consequently, it



(a) (b)

Figure 3.3: The path traversed by the space-filling curve based methods. (a) Hilbert, (b) Morton. Each patch is drawn with a different color (by using a colormap from blue to orange) where neighboring patches have similar colors. The black lines also connect the patches in the order they appear in the sequence.

is possible to use the space-filling curves to map a point from the 2-dimensional image space to the 1-dimensional sequence space. Moreover, certain space-filling curves provide this mapping in a way that also preserves the locality of the points, i.e. the points located in similar sections in the 1-dimensional sequence are also located in similar regions of the image space. When applied to our case, discrete approximations of the space filling curves can serve as the basis of an excellent sequence generating method due to their locality preservation quality. In this thesis, we conducted our experiments using two space-filling curves, Hilbert and Morton. Figure 3.3 illustrates the sequence generated by these methods.

The space-filling curves can be generated by following a recursive strategy. The process starts with a first order space-filling curve, which orders cells of a 2×2 grid. This first step is trivial as it is basically a reverse "u" shape for Hilbert and a "z" shape for Morton. To obtain the second order space-filling curve each cell is divided into 4 equal parts then ordered with a first order curve. Lastly, endpoints of the new curves are connected in accordance with the space-filling curve of the previous order. Even though the Hilbert curve requires some additional rotation operations, the higher order curves can be generated by re-applying these basic steps. Figure 3.4 provides an illustration of this process. Performing these steps

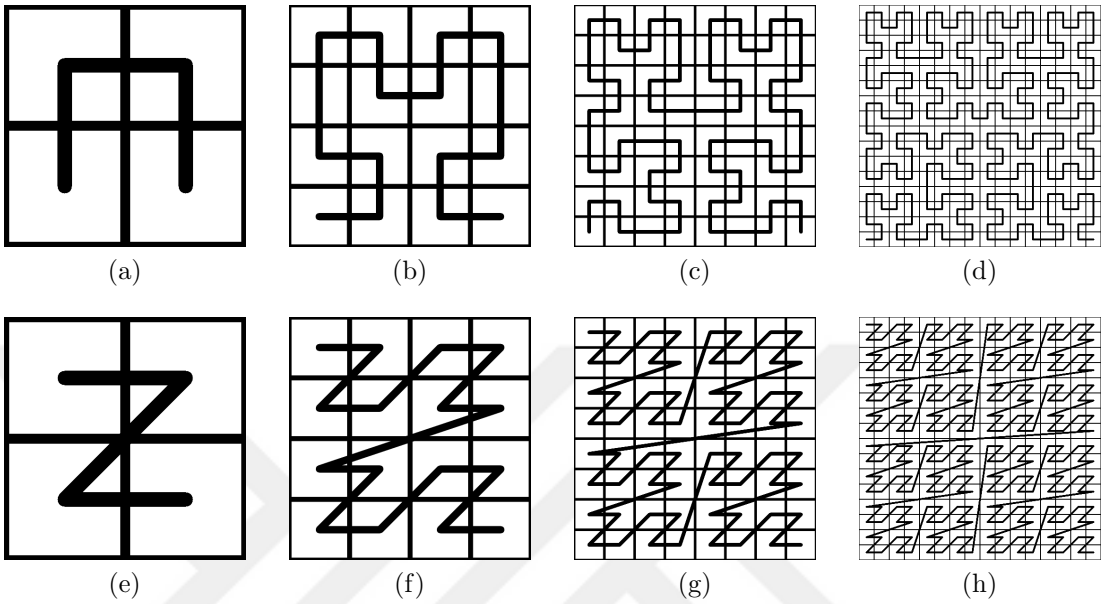


Figure 3.4: Generation of space-filling curves. (a)-(d) generation of Hilbert curve, (e)-(h) generation of Morton order.

infinitely many times would yield the space-filling curve in question, but a few repetitions are enough to order the patches of a WSI.

After obtaining a sufficiently high order approximation of a space-filling curve, the next step is ordering the foreground patches of the WSI according to the path described by the space-filling curve in question. To create such an order, we assign each patch an integer value that represents their location in the path. Then we remove background patches and sort the remaining foreground patches based on the values assigned. Then we use the output of this sorting step as the final patch sequence of the corresponding WSI.

Even though the space-filling curve based approach requires much more computational power compared to the simple methods, they also provide some significant advantages too. Firstly, locality preservation quality of the space-filling curves allows the model to safely assume that patch pairs that are close to each other in the sequence space are also reasonably close to each other in the WSI space. Because of this property of the space-filling curves, the model can recognize the structures even when they are composed of multiple patches. This

is significant, as some patch groups may be part of a diagnostically informative structure. Consequently, the ability to capture this information gives the transformer model an edge over the models that cannot understand the spatial context of the patches. Moreover, unlike the simple methods, space-filling curves are not limited to travel in a single axis, thus they are not oblivious to the patch neighbors in different axes. Considering the structures that constitute a WSI usually span multiple rows and columns of patches, this is an important advantage of space-filling curve based methods. Lastly, these methods usually handle the discontinuities in the foreground region with relative ease as they are not forced to follow a path in a single direction. When a simple method encounters a large hole in the foreground region, the sequence is required to continue on the other side of the discontinuous region. This causes the sequence to jump around the WSI and mark distant patches as neighbors. However, this is not the case for space-filling curve based methods. Even after discarding some patches, the remaining ones lie on a mostly continuous path. Consequently, when this path encounters an empty region, it does not cross it and the sequence continues in the same local region.

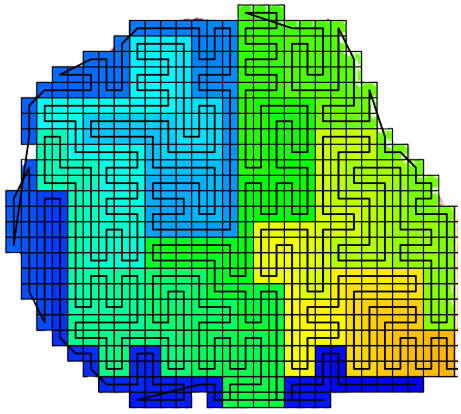


Figure 3.5: The path traversed by the content-based SFC method. Each patch is drawn with a different color (by using a colormap from blue to orange) where neighboring patches have similar colors. The black lines also connect the patches in the order they appear in the sequence.

3.2.3 Data Aware Sequences

The sequence generation methods discussed in this thesis so far work the same for all different slides. Because of that, they discard the underlying structure present in the WSI. However, a method that orders the patches while considering the actual structure of the slide can perform better by preserving the meaningful locality. Because of that, we experimented with a sequence generation method that is actually aware of the spatial distributions of the WSI patches. Figure 3.5 illustrates the sequence generated by this method.

In order to account for the actual distribution of the diagnostically relevant structures of the WSIs, in this thesis we experimented with the content based space-filling curves, proposed by Dafner et al. [30]. Even though the method proposed in this paper creates a path that traverses each pixel of the image, it can be modified to process the patched structure of a WSI too.

The process of generating a sequence is composed of three steps. In the first step, we divide the WSI into patch groups which are composed of 4 patches. We obtain the groups by following a non-overlapping sliding window approach with a 2×2 sized window. Then using these patch groups as vertices, we construct a directed and weighted graph. Each vertex is connected with an edge if the corresponding patch groups are adjacent to each other, either vertically or horizontally. We determine the edge weights using a cost function which quantifies the dissimilarity of two patch groups, i.e. the cost function yields a smaller value if two groups are similar on the RGB color space. In the second step, we calculate the minimum spanning tree corresponding to the graph we constructed in the first step using conventional graph processing algorithms. The resulting spanning tree provides some insight regarding the actual structure of the WSI. This is because the edge that connects patch groups has smaller value when two groups are similar. Consequently, only the similar looking patch groups remain as neighbors while the connection between other groups are severed. In the final step, we generate the patch sequence based on the minimum spanning tree calculated in the second step.

The most important advantage of this method is its ability to shape the sequence based on the WSI at hand. The resulting sequence aims to fully traverse each structure before traveling to another part of the image. Consequently, each consecutive patch of the sequence likely belongs to the same greater and possibly informative region of the slide. Moreover, unlike other methods discussed so far, with this method the holes in the foreground region should not create any discontinuous portions in the sequence, as the content based space-filling curve avoids them anyways. However, the most important drawback of this method is its inherently unpredictable nature. While other methods follow a strict path predefined by rules, this method can create arbitrarily shaped sequences. Consequently, it is possible that the sequence can be too random for the transformer model to understand.

3.3 Spatial Context Modeling

Generating the order of the patches is the first step of describing the spatial distribution of the patches. In order for the model to actually use this information, the order information should be supplied. However, since simply changing the order of the input tokens is meaningless, modeling the spatial context of the WSI requires slightly altering the feature vectors of the corresponding patches. This modification should be significant enough to communicate the order, but also it should not dominate the feature vectors as they are also the descriptors of the patches.

After determining the desired order of the patches, spatial context of the WSI can be modeled in many different ways. However, these methods can be clustered into two main groups. Methods in the first group communicate the position information in an absolute manner. Here, the input tokens are modified in a way that solely depends on their position in the sequence and nothing else. On the other hand, the methods of the second group focus on a more relative approach. Instead of providing the exact position of a token, they provide its position relative to other tokens. Rest of this section discusses these approaches and the

corresponding methods in detail.

3.3.1 Absolute Position Information

The spatial context modeling methods that transfer the absolute position information to the model aims to communicate the exact location of each patch in the input sequence. The main advantage of this method is it provides all available information to the model. Since it knows the exact location of each patch, technically it can deduce their relative locations too. The methods belonging to this group usually utilize special vectors called positional encodings to model the spatial context. Here, each location in the sequence is described by a unique positional encoding vector. These methods transfer the position information to the token by adding the corresponding position vector to its feature vector via an element-wise vector addition operation. However, it is important to note that directly adding the positional encoding vector can render the feature vector meaningless. To avoid that, only a scaled down version of the position vector is used in the addition operation.

While all methods follow a similar strategy to communicate the position information, the positional encoding vectors can be generated in multiple different ways. It is possible to separate these generation methods into two groups, learnable methods and constant methods. The learnable methods treat the positional encodings as learnable parameters of the model, and aim to learn the most suitable vectors during the training process. On the other hand, constant methods use a predefined positional encoding vector for each location of the token sequence.

Learnable Positional Encodings: The positional encodings for the learnable methods are initialized as random vectors and the model learns appropriate values for them as the training progresses. They are the most commonly used positional encoding method for the vision transformers as the learnable positional encodings can easily model the spatial context of an image with enough training. However, their usefulness significantly diminishes when used for WSI classification as the

tissue presented in the slide can be arbitrarily shaped. Moreover, this method requires a predefined limit on the maximum number of input tokens as each positional encoding vector should be initialized and trained. It is possible to circumvent this problem by setting a very high limit on the maximum number of input tokens. However, in this case the encodings that correspond to the later sections of the sequence would receive much less training as the number of patches within a WSI can be highly variable. It is also possible to cycle the position encodings to process more tokens, but in that case the model of the spatial context would be mostly meaningless. Because of these reasons, learnable positional encodings tokens are not a good fit for WSI classification. Consequently, in this thesis we focus on the constant positional encodings in the absolute case.

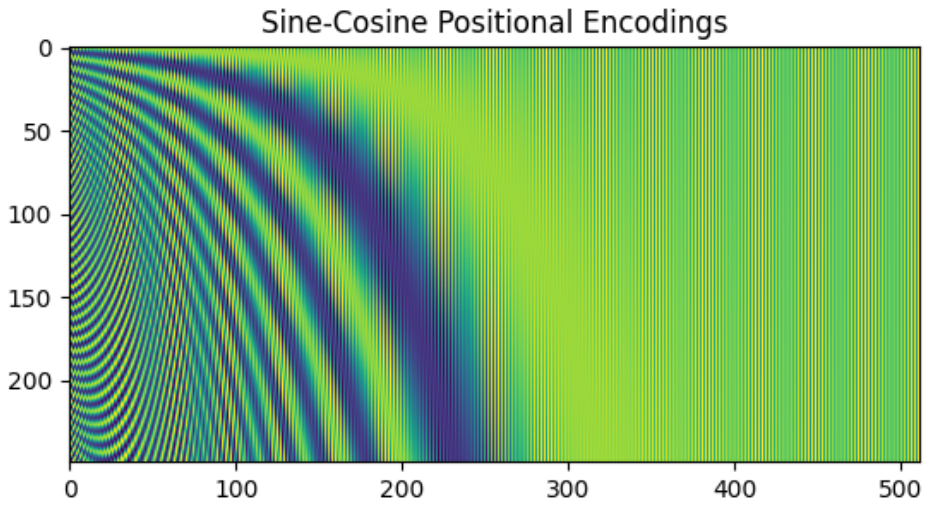


Figure 3.6: Heatmap of first 250 sine-cosine based positional encodings where yellow represents higher values. Each row represents the position encoding corresponding to that location in the sequence. Each column corresponds to a sine/cosine wave with a different frequency. Closely located rows yield a similar looking positional encoding.

Constant Positional Encodings: The methods that utilize constant positional encodings define a set of vectors before the training process begins. Each of these vectors corresponds to a unique location in the input sequence and they remain constant during the training phase. While any sufficiently large set of vectors can be used as positional encodings, in this thesis we used the sine-cosine based constant positions encodings proposed by Vaswani et al. [4]. The positional encodings for the n^{th} element of the sequence can be generated by the formula

$$PE\left[i\right] = \begin{cases} sin\left(\frac{n}{M^{i/d_{model}}}\right), & \text{i is even;} \\ cos\left(\frac{n}{M^{i/d_{model}}}\right), & \text{otherwise.} \end{cases} \tag{3.2}$$

where i is an index of the positional encoding vector, M is a value that is proportional to the maximum number of vectors and d_{model} is the size of the feature vectors.

The sine-cosine based position encodings model the spatial context by utilizing sine and cosine waves with different frequencies. The waves with a high frequency provide a zoomed-in view of the sequence. This helps the model to distinguish patches that are located in close proximity. However, they do not provide any details regarding patch's actual position within the sequence. On the other hand, low frequency waves provide a zoomed-out view, pointing to the patch's general location in the whole sequence while supplying very little information regarding its position relative to its neighbors.

The most important advantage of the sine-cosine based encodings is that they are relatively unaffected by the varying sequence lengths. Even though they too limit the maximum sequence length, unlike learnable encodings, picking a high limit does not significantly reduce their performance as they do not require any training data. Moreover, when the sequence exceeds the maximum length, the positional encodings do not become completely meaningless due to the nature

of sinusoidal waves. Even though the constant nature of these positional encodings reduces the learning capacity of the transformer based classifier, they are a suitable position encoding method for WSI classification.

The sine-cosine encodings paint a pretty good picture of the order of the sequence. However, they do not provide any learning opportunities to the model and they inherently work on 1-dimensional sequences. To this end, in this thesis we propose three improvements over the original implementation of the sine-cosine based constant positional encodings. The first two methods, namely "Trainable Constant" and "Frequency Learning", aim to increase the learning capacity of the model by adding learnable sections to the spatial context modeling process. The third method aims to model the spatial context of the WSI in two dimensions, instead of ordering the patches in a 1-dimensional manner.

Trainable Constant: As mentioned previously, the main issue of the constant positional encodings is their inability to evolve based on the needs of the model. Consequently, the model has to work with what is given, reducing its learning capacity. As a solution to this problem, we experimented with a method proposed by Devlin et al. [15] that allows the positional encodings to change based on the requirements of the task at hand. The Trainable Constant method works by introducing a fully connected layer between the positional encodings and tokens. Instead of adding the positional encodings directly to the feature vectors, we first pass them through a trainable layer and use the output of this layer as the positional encodings. This approach allows the model to shape the positional encodings based on its needs. Moreover, some sections of the feature vectors might be too significant to be altered by the positional encodings. This method allows the model to reorganize the positional encoding vector in a way that preserves certain portions of the token descriptors.

Frequency Learning: The sinusoidal waves with different frequencies that constitute the sine-cosine encodings provide a view of the sequence at different zoom levels. Since the relative importance of different zoom levels are not known,

these encodings give the same priority to each different frequency. However, it might be possible that the information provided by some of these frequencies are relatively more significant. Nevertheless, the constant positional encodings do not allow the model to favor any frequency over others. To eliminate this limitation, we propose using Frequency Learning proposed by Wang et al. [31]. We achieve that by replacing the constant frequencies of the sine-cosine encodings with learnable parameters. Consequently, instead of using the same positional encodings for all tasks, this method allows the model to favor some frequencies over others. Furthermore, with this strategy the model can even use frequencies that are not present in the original form of the sine-cosine based positional encodings.

2-Dimensional Positional Encodings: All the methods discussed so far aim to convert an image into a sequence that could be understood by a transformer based classifier in a way that preserves the spatial context of the image. While most transformer based methods follow a similar approach, with slight modifications the sine-cosine based positional encodings can directly encode the 2-dimensional structure of an image as well. The proposed method achieves that by encoding X and Y coordinates of a patch, instead of its position in the sequence. This can be done by splitting the positional encoding vector into two parts, and using the first part to encode the X coordinate and the second part to encode the Y coordinate. Since these encodings are already designed to encode integers, the only modification necessary is creating two half-sized vectors and concatenating them to obtain the vector with appropriate size. In theory, unlike sequence based methods, this approach can perfectly model the spatial context of a WSI without losing any information.

3.3.2 Relative Position Information

The methods that model the spatial context in an absolute manner aim to clearly define the position of each patch in the sequence. However, since the main premise of the transformers is calculating the pairwise attention of the patches, providing

only the relative position information might be sufficient as well. Because of that, the methods that provide the relative position information, only communicate the positions of the patches relative to each other.

Relative position information can be supplied to the model in different ways, however not all methods are applicable to our case. For example, the relative position information model proposed by Huang et al. [32] works by altering the values of the QK^T matrix. This approach changes the attention scores of the tokens based on their relative locations. However, dealing with the QK^T matrix is a very costly process in terms of memory and computational power. Considering the cost increases quadratically, this is especially problematic for WSIs, as the sequences can be very long. In order to avoid dealing with such high costs, in this thesis we propose employing a convolution based relative position information model.

Table 3.1: Details of the proposed convolution based methods for the relative position information approach. The subscripts RC, LN and GC stand for the use of residual connection, layer normalization and group convolution, respectively.

	Residual Connection	Layer Normalization	Group Convolution
$\text{conv}_{\text{base}}$			
conv_{LN}		✓	
conv_{RC}	✓		
$\text{conv}_{\text{RC,LN}}$	✓	✓	
conv_{GC}			✓
$\text{conv}_{\text{RC,GC}}$	✓		✓

The proposed convolution based spatial context modeling method employs 1-dimensional convolution operations to communicate the spatial context of the image. In order to apply 1-dimensional convolution, we treat the patches as a sequence with 512 channels. The method uses 3 convolutional layers with kernel sizes 3, 5 and 7 with no stride or dilation. We pad each feature vector with the appropriate values to keep the number of input tokens constant as normally the convolution operation reduces the length of the sequence. After applying the convolutions, we obtained 3 new sequences. Then, we summed all 3 outputs to obtain a new set of tokens.

In this thesis, we experiment with 6 convolutional architectures that differ in terms of the use of residual connection, layer normalization and group convolution. Table 3.1 lists the details of these convolutional models that are named depending on the use of these features. Overall, 3 of the models use residual connection, 2 of them employ layer normalization and 2 of them utilize group convolution. Convolutional models are different from the models presented so far in this thesis in the way they model the spatial context. Unlike prior model architectures that model the spatial context by adding positional encodings to feature vectors, convolutional architectures model the spatial context implicitly by ordering the feature vectors of the patches. Convolutional methods utilize the order of the sequences as follows: 1) they order the patches based on the generated sequence so that the neighboring patches with respect to these sequences are provided to the model sequentially, 2) apply convolution to the neighboring patches together by adapting the kernel size. In this architecture, a kernel size of 5 results in applying a convolution to a patch together with the two patches that follow the patch and precede the patch jointly, providing the spatial context to the model implicitly. Even though this approach models the spatial context, thus the relationship of neighboring patches quite well, it does not provide any positional information regarding the non-neighbors of a patch. Since this method does not employ positional encodings, this method does not exactly follow the model overview presented in Figure 3.1.

Chapter 4

Experiments and Results

This chapter provides the details regarding the datasets, experimental setting and conducted experiments and their results.

4.1 Datasets

To train and test the proposed methods, we used the lung cancer WSIs released by The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) under projects TCGA-LUAD and TCGA-LUSC. Our dataset contains 532 LUAD (Lung Adenocarcinoma) slides and 508 LUSC (Lung Squamous Cell Carcinoma) slides. We conducted our experiments in a binary setting, where the model aims to differentiate between these two types of lung cancer.

The slides in our dataset have an average size of $90,000 \times 62,000$ pixels. From those slides, we extract about 26,000 patches. During the feature extraction process, we discard approximately half of these patches as they correspond to the background regions of the WSI.

In order to use the dataset in our experiments, we split the dataset into 5 folds. Each fold contained roughly the same amount of samples. In order to avoid data

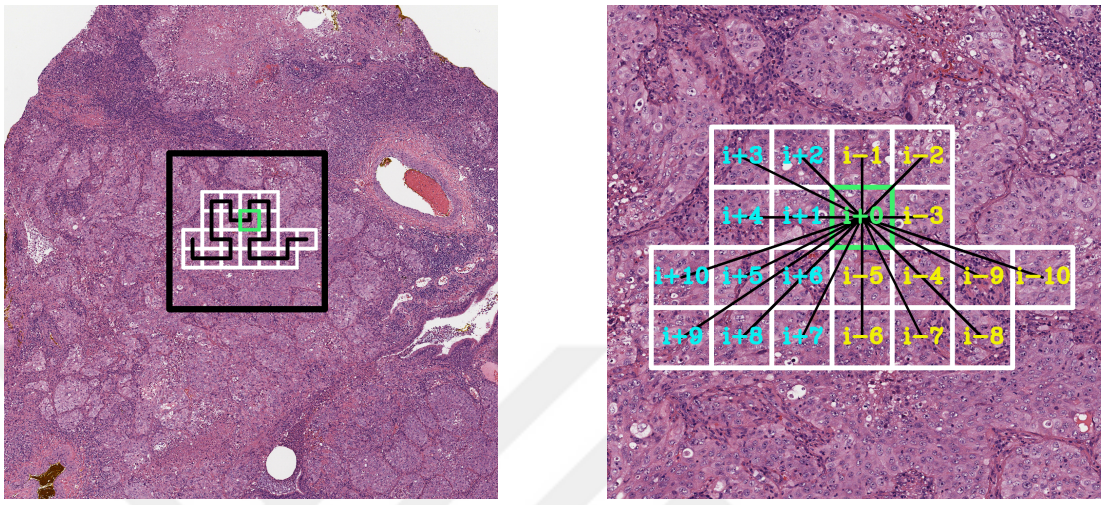
leaks between folds, slides obtained from the same patients remained in the same fold.

4.2 Experimental Setting

We trained and tested our models in two steps. In the first step, we used 3 folds for training and 1 fold for validation. We used this step to tune the hyperparameters of the model. Then we fixed the best performing hyperparameters and moved on to the second step. In the second step, we used 3 folds for training, 1 fold for validation and 1 fold for testing. In this step, we did not perform any hyperparameter tuning, and used the validation set only to pick the best performing model. Then, we cycled the training, validation and test folds 5 times to conduct 5 different experiments for each method. This allowed us to use every single sample in the dataset for testing exactly once and improved the confidence of the performance results.

4.3 Analysis of Sequence Generation Methods

As mentioned previously, we are hypothesizing that an appropriate sequence should maintain the spatial integrity of regions by preserving the locality of the patches. Moreover, it should be robust to the variations of the shape and size of the WSIs while handling the possible discontinuities in the foreground region. In order to quantify the locality preserving quality of different ordering methods, we came up with a metric called Mean Distance to N Nearest Neighbors (MDNNN), which measures a patch's mean distance to its sequence neighbors in the slide space. Figure 4.1 illustrates MDNNN calculation for a single patch. We calculate MDNNN score for a single slide by averaging MDNNN scores of all of its patches. Then, to obtain a dataset-wise MDNNN score, we average slide MDNNN scores for all slides in the dataset.



(a) (b)

Figure 4.1: (a) An example patch and its neighbors in the sequence space. (b) Illustration of MDNNN calculation for the i-th patch in the Hilbert sequence for $N = 2 \times 10$. In this case, MDNNN is calculated as the average of the i-th patch's distance in the 2D slide space to each of the 10 predecessor and 10 successor patches in the 1D sequence.

Figure 4.2 presents MDNNN scores for different N values for different sequence generation methods. Results show that Hilbert and content based space-filling curve methods perform significantly better than other sequence generation methods for all N values, meaning they preserve the locality better than other methods. Performance of the Morton increases as N increases compared to spiral and scan. However for smaller N values, all three perform similarly. Spiral performs slightly better than scan, but their MDNNN scores are fairly similar for every N value.

4.4 Classification Experiments

In this section, we present the classification experiments we conducted. The first part of this section covers the experiments related to the sequence generation methods. In the second part we present the experiments regarding the discussed spatial context modeling methods. The final part provides other experimental results for the patch size and scaling factor of positional encodings.

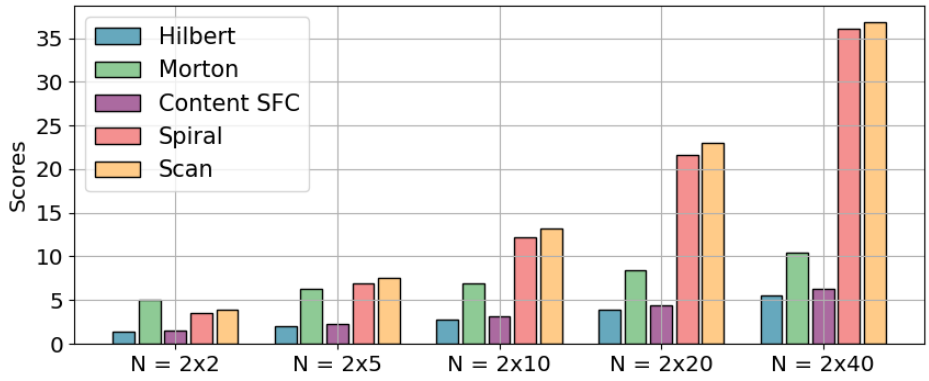


Figure 4.2: MDNNN scores of different sequence generation methods for different N values. Smaller values indicate better performance.

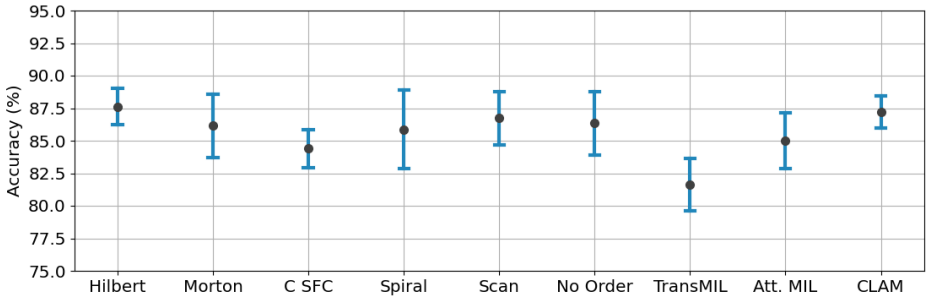


Figure 4.3: Classification accuracies of the sequence generation methods.

4.4.1 Sequence Generation Experiments

To determine the best performing sequence generation method, we performed the WSI classification experiment using all proposed methods. To model the spatial context of the image, we used the sine-cosine based spatial context modeling method described in Section 3.2 as it is the most commonly used approach. We also compared our results with the case where no position information was provided (No Order), as well as the state-of-the-art WSI classification networks. Among these methods, CLAM and Attention MIL are multiple instance learning based classification models. On the other hand, TransMIL is a transformer-based WSI classification framework. The results of the experiment for each test fold are given in Table 4.1. Figure 4.3 presents the average classification performance

with error bars denoting 95% confidence intervals.

Table 4.1: Classification accuracies for the sequence generation experiment for each test fold.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Avg ± Stdv**
Hilbert	86.67%	87.62%	86.89%	86.47%	90.34%	87.6 ± 1.59
Morton	85.71%	88.57%	88.35%	81.64%	86.47%	86.1 ± 2.8
Content SFC	87.08%	83.81%	84.39%	82.61%	84.06%	84.4 ± 1.65
Spiral	87.62%	88.10%	84.95%	80.19%	88.41%	85.9 ± 3.45
Scan	85.71%	87.62%	86.89%	83.57%	89.86%	86.7 ± 2.33
No Order	83.81%	89.05%	86.89%	83.09%	88.89%	86.3 ± 2.79
TransMIL [20]	80.95%	78.10%	83.01%	82.13%	84.06%	81.7 ± 2.29
Att. MIL [13]	80.95%	85.71%	85.92%	85.02%	87.44%	85.0 ± 2.43
CLAM [14]	89.05%	86.19%	86.41%	85.99%	88.41%	87.2 ± 1.41

The method that utilizes Hilbert space-filling curves to generate the sequence, performs best in terms of mean accuracy by achieving 86.7% accuracy on average, even surpassing other state-of-the-art WSI classification frameworks. Considering Hilbert curve also performs best in terms of its MDNNN score, its performance also supports the idea that the sequence generation methods that preserve the locality of the patches provides a better strategy for modeling the spatial context of a WSI. On the other hand, content based space-filling curve's performance suggests otherwise as it is the worst performing method even though it achieves an MDNNN score comparable to the Hilbert. However, this might be due to its inherently unpredictable approach to sequence generation as all other methods generate the sequences in a somewhat predictable manner.

Among other sequence generation methods, the raster scan order achieves the best performance with an average accuracy of 86.7%, followed by Morton and spiral orders, respectively. However, all three of these models perform similar to the case where no information regarding the spatial context of the WSI is provided. This result might mean that the information they provide is not useful

for the classification model at all. Note that these methods also perform significantly worse than Hilbert curve based method in terms of their MDNNN score. Consequently, it might be concluded that in order to model the spatial context of a WSI successfully, a sequence generation method should have a lower MDNNN score.

We also conducted McNemar tests to assess the significance of the performance improvements. According to the results of these tests, Hilbert curve based method performed statistically better than all other methods; except for CLAM, scan and no order with 95% confidence.

4.4.2 Spatial Context Modeling Experiments

Our first experiment regarding spatial context modeling is the Trainable Constant method. Since this method requires an order to be generated first, we experimented with three different order generation methods. In order to evaluate the performance of each sequence generation approach, we conducted our experiments with Hilbert curve, scan and content based space-filling curve methods. Results of these experiments are given in Table 4.2. In this table, the preceding "T" denotes the Trainable Constant method.

Table 4.2: Experiment results for the Trainable Constant spatial context modeling method.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average ± Stdv**
THilbert	81.90%	83.33%	81.07%	76.8%	85.02%	81.6 ± 3.08
TScan	82.86%	83.33%	79.61%	82.13%	81.16%	81.8 ± 1.48
TCSFC	83.73%	81.90%	81.95%	79.71%	85.02%	82.5 ± 2.02

The results show that each sequence generation method performs significantly worse when the spatial context is modeled by the Trainable Constant method. Performance of the Hilbert and scan methods drop about 6%, while the performance of the content based SFC model drops 1.9%. According to our results, the

trainable layer added after the position embeddings does not allow the model to understand spatial context better.

Next, we experimented with the method that allows us to model the spatial context in 2-dimensions instead of creating a sequence of patches. Since this approach does not require an order to be generated, we did not perform any additional experiments for different order generation methods. However, the structure of this method allows it to be used in conjunction with the Trainable Constant method, so we experimented with the combination of these two approaches as well. Results of this experiment are given in Table 4.3, again the preceding "T" stands for the Trainable Constant method.

Table 4.3: Experiment results for the 2-dimensional spatial context modeling method.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average ± Stdv**
2Dsine	85.71%	86.19%	85.92%	82.13%	82.61%	84.5 ± 1.97
T2Dsine	85.71%	84.29%	80.58%	82.61%	86.47%	83.9 ±2.38

Even though the 2-dimensional modeling method theoretically perfectly models the spatial context, experimental results show that it does not perform better than the sequence based approaches. The results suggest that this approach might be too complicated for the transformer model. Note that the sine-cosine based approach works by slightly altering the input tokens of the transformer. While it is possible to communicate the 1-dimensional position information by slightly nudging the tokens, it becomes an indecipherable mess in the 2-dimensional case. Since this approach also changes the input tokens, it underperforms even when compared to the case where no position information is provided. Similar to the previous experiment, the Trainable Constant method again negatively affects the performance of the classifier.

As the final experiment of the absolute position information approach, we tried to experiment with the Frequency Learning method. However, since the model aims to learn the ideal frequencies during the training process, the positional encodings should be generated from scratch every time the model weights

are updated. Considering the number of patches extracted from each WSI, this approach becomes excessively costly in terms of computational power. Consequently, this method is not a good fit for WSI classification. Moreover, our initial findings suggest that it will not perform better than our methods, even when trained fully.

To evaluate the performance of the methods that provide the position information in a relative manner, we conducted experiments with the convolutional methods discussed in the section 3.3. We used the Hilbert curve based sequence generation method to create patch sequences since it performed better than other sequence generation methods. The results of this experiment are presented in Table 4.4 whereas the exact details of the convolution models are given in Table 3.1.

Table 4.4: Experiment results for the convolution based spatial context modeling methods. The subscripts RC, LN and GC stand for the use of residual connection, layer normalization and group convolution, respectively.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average ± Stdv**
$\mathbf{conv_{base}}$	85.24%	85.71%	88.35%	81.16%	87.44%	85.6 ± 2.77
$\mathbf{conv_{LN}}$	86.19%	81.90%	84.47%	81.64%	87.44%	84.3 ± 2.56
$\mathbf{conv_{RC}}$	85.24%	86.67%	86.89%	79.71%	89.37%	85.6 ± 3.60
$\mathbf{conv_{RC,LN}}$	83.33%	83.33%	85.92%	78.74%	84.06%	83.1 ± 2.65
$\mathbf{conv_{GC}}$	86.19%	84.76%	84.47%	81.16%	86.47%	84.6 ± 2.12
$\mathbf{conv_{RC,GC}}$	85.71%	85.24%	84.47%	82.13%	85.02%	84.5 ± 1.41

According to our results, modeling the spatial context in a relative manner with a convolution based approach does not introduce any performance improvements. Among the architectures we tested, the models which did not utilize neither layer normalization nor group convolution performed the best. On the other hand, using residual connections after the convolution step did not have any noticeable impact on the average classification accuracy. The results might suggest that the kernel size we used caused the model to be oblivious to most of its neighbors as it was too small compared to the length of the input sequence.

4.4.3 Other Experiments

In this section, we lay out the experiments we conducted to determine the optimal scaling factor of the positional encodings as well as the ideal patch size for the feature extraction process.

As mentioned previously, the positional encodings are added to the input tokens to model the spatial context. However, they are scaled down to preserve the meaning of the original input features. Here, determining the optimal scaling factor is crucial. Because a scaling factor that is too small might not be able to communicate the spatial context, while it might dominate the feature vectors otherwise. To this end, we performed two more experiments with higher and lower values of the scaling factor. In those experiments we used the Hilbert based method to generate the sequence and sine-cosine based positional encodings to model the spatial context. Results of these experiments are given in Table 4.5.

Table 4.5: Experiment results for different scaling factors.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average ± Stdv**
0.1	61.90%	64.29%	54.37%	51.69%	70.53%	60.6 ± 7.62
0.01	86.67%	87.62%	86.89%	86.47%	90.34%	87.6 ± 1.59
0.001	87.62%	85.24%	83.01%	79.23%	86.47%	84.3 ± 3.32

Table 4.6: Experiment results for different patch sizes.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average ± Stdv**
256×256	86.67%	87.62%	86.89%	86.47%	90.34%	87.6 ± 1.59
512×512	86.32%	86.60%	83.33%	80.48%	84.29%	84.2 ± 2.49

For the experiments we discussed previously, we used 0.01 as the scaling factor. The results show that using higher or lower values negatively affects the performance of the proposed classification network.

Finally, we conducted an experiment to determine the ideal patch size. Even though the proposed model can work with any patch size, different patch sizes

have different advantages and disadvantages. For example, using a larger patch size results in a shorter sequence and makes it easier to model the spatial context. However, it makes the feature extraction step prone to errors as the feature extraction model might miss smaller details of a larger patch. On the other hand, using smaller sized patches quadruples the sequence length and thus significantly increases the training time while making it harder to model the spatial context. In our experiments we used 256×256 sized patches, but we also experimented with a model that utilizes 512×512 sized patches. Results of this experiment are presented in Table 4.6.The results show that performing the WSI classification with larger patches negatively affects the performance of the classification model.

Chapter 5

Conclusions and Future Work

Most WSI classification frameworks follow a patch based strategy to overcome the challenges introduced by the enormous sizes of WSIs. They extract patch features and combine the patch representations to classify WSIs. It is possible to perform this operation with a transformer-based classifier as well. However, since transformers cannot understand the distribution of the patches in the two-dimensional space, such an approach also requires a method of modeling the spatial context of the patches. This is a challenging task as conventional methods only provide an informative representation of the image only when its size is fixed. In this thesis, we proposed a novel spatial context modeling strategy in order to improve the performance of the classifier by providing an accurate view of patch locations.

Providing the position information to the transformer requires two steps. First the patches should be ordered as a sequence, and then this order information should be transferred to the model. The generated sequence should be adaptable to the variations of the tissues in the WSI in order to successfully model the locations of the patches. To this end, we proposed using two potentially promising sequence generation approaches. The first approach uses space-filling curves to exploit their ability to preserve the locality of the patches. The second method aims to generate a sequence that is capable of understanding the actual structure

of WSI by using content based space-filling curves.

For the second step, we experimented with different methods of supplying the order information to the transformer model. To present the absolute position information, we used the commonly used sine-cosine based position encoding method as well as some of its variants. Since this information can also be provided in a relative manner, we also experimented with a convolution based relative position information scheme.

In our experiments, Hilbert curve based sequence generation method performed better than state-of-the-art WSI classifiers on average when used with sine-cosine based positional encodings. However, other sequence generation methods did not provide any performance improvements.

In the future work, the proposed relative position information approach can be improved by using different CNN architectures. In this thesis, we experimented with only two space-filling curves. Thus the performance of other space filling-curves can be investigated. Lastly, the performance of the model can be improved by utilizing WSI region of interest detection tools.

Bibliography

[1] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-Instance Multi-Label Learning for Multi-Class Classification of Whole Slide Breast Histopathology Images," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 316–325, Jan. 2018.

[2] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized Classification of Intraductal Breast Lesions Using Histopathological Images," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 1977–1984, July 2011.

[3] E. Cosatto, P.-F. Laquerre, C. Malon, H.-P. Graf, A. Saito, T. Kiyuna, A. Marugame, and K. Kamijo, "Automated Gastric Cancer Diagnosis on H&E-Stained Sections; Training a Classifier on a Large Scale With Multiple Instance Machine Learning," in *Medical Imaging 2013: Digital Pathology* (M. N. Gurcan and A. Madabhushi, eds.), vol. 8676 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 867605, Mar. 2013.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," 2017. arXiv:1706.03762.

[5] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.

[6] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," 2020. arXiv:2004.05150.

[7] Y. Xiong *et al.*, "Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[8] C. Erkan and S. Aksoy, "Space-filling curves for modeling spatial context in transformer-based whole slide image classification," in *Medical Imaging 2023: Digital and Computational Pathology* (J. E. Tomaszewski and A. D. Ward, eds.), vol. 12471, p. 124711L, International Society for Optics and Photonics, SPIE, 2023.

[9] Y. Zheng, Z. Jiang, F. Xie, J. Shi, H. Zhang, J. Huai, M. Cao, and X. Yang, "Diagnostic Regions Attention Network (DRA-Net) for Histopathology WSI Recommendation and Retrieval," *IEEE Transactions on Medical Imaging*, vol. 40, pp. 1090–1103, Mar. 2021.

[10] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images," *Nature Medicine*, vol. 25, pp. 1301–1309, Aug. 2019.

[11] B. Xu, J. Liu, X. Hou, B. Liu, J. Garibaldi, I. O. Ellis, A. Green, L. Shen, and G. Qiu, "Look, Investigate, and Classify: A Deep Hybrid Attention Method for Breast Cancer Classification," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 914–918, 2019.

[12] T. Vu, P. Lai, R. Raich, A. Pham, X. Z. Fern, and U. A. Rao, "A Novel Attribute-Based Symmetric Multiple Instance Learning for Histopathological Image Analysis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3125–3136, 2020.

[13] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-Based Deep Multiple Instance Learning," 2018. arXiv:1802.04712.

[14] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images," *Nature Biomedical Engineering*, vol. 5, pp. 555–570, June 2021.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. arXiv:1810.04805.

[16] D. Wang, W. Xie, Y. Cai, X. Li, and X. Liu, "Transformer-Based Rapid Human Pose Estimation Network," *Computers & Graphics*, 2023.

[17] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse Self-Attention Transformer for Image Inpainting," *Pattern Recognition*, p. 109897, 2023.

[18] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo, A. Smargiassi, G. Soldati, P. Rota, A. Passerini, R. J. G. van Sloun, E. Ricci, and L. Demi, "Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2676–2687, 2020.

[19] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, 2022.

[20] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, *et al.*, "TransMIL: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021.

[21] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images," pp. 3995–4005, 2021.

[22] A. G. Gul, O. Cetin, C. Reich, N. Flinner, T. Prangemeier, and H. Koeppl, "Histopathological Image Classification Based on Self-Supervised Vision Transformer and Weak Labels," in *Medical Imaging 2022: Digital and Computational Pathology* (J. E. Tomaszewski, A. D. Ward, and R. M. L. M.D., eds.), vol. 12039, p. 120391O, International Society for Optics and Photonics, SPIE, 2022.

[23] S. Mehta, X. Lu, W. Wu, D. Weaver, H. Hajishirzi, J. G. Elmore, and L. G. Shapiro, "End-to-End Diagnosis of Breast Biopsy Images With Transformers," *Medical Image Analysis*, vol. 79, p. 102466, 2022.

[24] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-Attention With Relative Position Representations," 2018. arXiv:1803.02155.

[25] T. Asano, D. Ranjan, T. Roos, E. Welzl, and P. Widmayer, "Space-Filling Curves and Their Use in the Design of Geometric Data Structures," *Theoretical Computer Science*, vol. 181, no. 1, pp. 3–15, 1997.

[26] F. Li, R. Chen, C. Zhou, and M. Zhang, "A novel Geo-Spatial Image Storage Method Based on Hilbert Space Filling Curves," in *2010 18th International Conference on Geoinformatics*, pp. 1–4, 2010.

[27] H.-L. Chen and Y.-I. Chang, "Neighbor-Finding Based on Space-Filling Curves," *Information Systems*, vol. 30, no. 3, pp. 205–226, 2005.

[28] S. Biswas, "One-dimensional B-B Polynomial and Hilbert Scan for Graylevel Image Coding," *Pattern Recognit.*, vol. 37, pp. 789–800, 2004.

[29] S. Anders, "Visualization of Genomic Data With the Hilbert Curve," *Bioinformatics*, vol. 25, pp. 1231–1235, Mar. 2009.

[30] R. Dafner, D. Cohen-Or, and Y. Matias, "Context-based Space Filling Curves," *Computer Graphics Forum*, vol. 19, 05 2000.

[31] B. Wang, L. Shang, C. Lioma, X. Jiang, H. Yang, Q. Liu, and J. G. Simonsen, "On Position Embeddings in BERT," in *International Conference on Learning Representations*, 2021.

[32] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, and D. Eck, "Music Transformer: Generating Music with Long-Term Structure," 2018. arXiv:1809.04281.