Order-ViT: Order Learning Vision Transformer for Cancer Classification in Pathology Images

Ju Cheon Lee and Jin Tae Kwak
School of Electrical Engineering, Korea University, Seoul, Republic of Korea
dlwncjs0618@korea.ac.kr, jkwak@korea.ac.kr

Abstract

In computational pathology, cancer classification is one of the most widely studied tasks. There exist numerous tools for cancer classification, which are mainly built based upon convolutional neural networks or Transformers. These tools, by and large, formulate cancer classification as a categorical classification problem, which ignores the intrinsic relationship among cancer grades. Herein, we propose an order learning vision transformer for cancer classification that can not only learn the histopathological patterns of individual cancer grades but also utilize the ordering relationship among cancer grades. Built based upon vision transformer, the proposed method simultaneously conducts categorical classification per input sample and order classification for a pair of input and reference samples. Moreover, it introduces a voting scheme to identify less confident samples and to improve the accuracy of the decision on such samples. The proposed method is evaluated on two types of cancer datasets including colorectal and gastric cancers. Experimental results show that the proposed method outperforms other classification models and can facilitate improved cancer diagnosis in clinics.

1. Introduction

Cancer is one of the leading causes of death that affects health worldwide [3]. Rapid and accurate diagnosis of cancer is critical to initiating timely treatment and improving patient outcomes. The current gold standard for cancer diagnosis involves the microscopic examination of tissue samples stained with hematoxylin and eosin (H&E) by pathologists. However, these manual histopathological evaluations are time-consuming, low-throughput, and prone to inter- and intra-observer variability, and thus can limit the efficiency and accuracy of cancer diagnosis. Therefore, an automated, fast, and reliable methods are needed to improve the quality of cancer diagnosis and patient care.

Computational pathology, which combines artificial intelligence (AI), whole-slide imaging (WSI), and clinical informatics, is an emerging field of study that hold great potential for transforming and improving the practice of current pathology [2, 4, 8]. Many of these tools rely on deep learning algorithms, especially convolutional neural networks (CNNs). CNNs have been successfully applied to numerous applications including cancer diagnosis [43], tissue segmentation [16], and synthetic tissue image generation [38]. Recently, Vision Transformer (ViT) has received considerable attention for its excellent performance in computer vision tasks such as object detection [5], semantic segmentation [30], and scene understanding [40]. The application of ViT has been also extended to pathological image analysis such as histopathology image classification [10, 34], metastasis detection [4], and survival analysis [23].

For both CNN-based and ViT-based models, cancer diagnosis has been primarily studied as a categorical classification problem. As a categorical classification problem, the role of CNNs or ViT is to identify the unique and distinct patterns of each cancer grade within a tissue image. In this perspective, Cancer grades are considered to be independent to each other. Although this approach has been successful, it, by and large, ignores the intrinsic relationships among cancer grades. It is generally accepted that the higher the cancer grade, the more aggressive it is. This indicates that there is an intrinsic ordering among cancer grades. [18], in fact, pointed this out and proposed a CNN-based model that conducts cancer grading as both categorical classification and regression problems, which handles one tissue image at a time. Moreover, the ordering relationship can be incorporated into cancer diagnosis via pairwise comparison between tissue samples. Suppose that there are borderline cases between two cancer grades such as high-grade cancer and low-grade cancer, potentially leading to mis-diagnosis in the clinics. For such cases, it will be easier or helpful to make a comparison against some representative cases of the two grades and tell whether is less or more aggressive than the representative cases. This approach of learning the ordering relationship is called order learning [20].

In this work, we propose an order learning ViT (Order-
ViT) for cancer grading in pathology images. Order-ViT conducts two classification tasks simultaneously: one for categorical classification and the other for order classification. For the categorical classification, it receives one tissue sample and predicts its class label, i.e., learning and exploiting specific histopathological patterns of cancer grades. As for order classification, it takes an additional reference tissue sample and predicts the ordering relationship between the two sample via pairwise comparison. Order-ViT contains a feature extractor, a categorical classifier, and an order classifier. We employ both CNN and ViT to build a feature extractor that extracts the high-dimensional feature representation of a tissue sample. Subsequently, the categorical classifier assigns a cancer grade per tissue sample and the order classifier receives the feature representations of a pair of tissue samples and predicts the ordering relationship between them. The categorical classifier conducts the primary classification. If and only if the result of the primary classification is uncertain, we invoke a voting scheme that conducts a number of ordinal classifications and voting to determine the final prediction.

Our contributions are summarized as follows:

- We propose an order learning ViT for cancer grading in pathology images, which learns both the histopathological patterns and the ordering relationship among cancer grades.
- We introduce a voting scheme that exploits the ordering relationship between tissue samples to further improve the accuracy and reliability of cancer grading.
- We evaluate the proposed method on two types of cancer datasets and compared it with CNN- and ViT-based models, outperforming other models regardless of the datasets.

2. Related Work

2.1. Computational Pathology for Cancer Diagnosis

A wide range of approaches and methods in computational pathology have been applied to cancer diagnosis. Early works focused on extracting and utilizing hand-crafted features that are associated with the presence and aggressiveness of cancer. The hand-crafted features include color intensity/histograms, morphological features, and texture features. Morphological features are to quantify the shape, distribution, and density of histological objects such as glands [29] and nuclei [14]. Texture features include grey-level co-occurrence matrix [27], wavelet transform [6], local binary pattern [25], and Gabor filters [24]. Later, deep learning methods, in particular CNNs, have shown to be effective in processing and analyzing pathology images. Various kinds of CNN architectures such as AlexNet [17], ResNet [9], EfficientNet [31], and MobileNet [11], which were developed for computer vision tasks, have been successfully applied to cancer diagnosis for colon cancer [26], gastric cancer [39], and breast cancer. In addition to these, a number of advanced models that are specific to pathology image analysis have been proposed. For example, [33] proposed a multi-scale CNN that extracts and utilizes binary patterns of feature maps across multiple scales for colorectal and prostate cancer classification. [19] adopted dual-domain attention mechanism to strengthen feature representations and to focus on important areas, achieving improved classification performance on colon and lung cancers. [42] introduced attention high-order model that simultaneously integrates the high-order statistics and attention mechanism modules into a residual network for breast cancer classification. Recently, Transformer has been adopted for several computer vision tasks as well as pathology image classification tasks, outperforming CNN-based models.

2.2. Vision Transformer (ViT)

Transformer was first introduced in the machine translation task in natural language processing. Vision Transformer (ViT), which is a variant of Transformer, was developed for the image classification task. ViT unfolds an input image as a series of small patches, i.e., tokens, so that Transformer layers can be utilized to incorporate the global context of the input image and to conduct the image classification task. This approach has been successfully adopted for various computer vision tasks, including image segmentation [15], image generation [37], and change detection of a pair of co-registered images [1]. It has been also widely applied to pathology image analysis. For example, [13] employed ViT for prostate cancer grading in WSIs by extracting and utilizing multiple image patches from WSIs. [41] conducted tissue segmentation in colorectal cancer images by adopting Swin Transformer for extracting the global context of tumour-related regions and a cascaded upsampler for detecting the tumor boundary. [35] proposed a hybrid architecture of Swin transformer and CNN that is equipped with contrastive learning to enhance feature extraction and to conduct five downstream tasks such as patch retrieval, patch classification, WSI classification, mitosis detection, and colorectal adenocarcinoma gland segmentation. [21] also exploited both CNN and ViT models to extract features from pathological images and then combine them to perform the classification of cervical cancer images.

2.3. Order Learning

Order learning is a learning paradigm that aims at learning the ordered relationship among class labels, which was first introduced in age estimating [20]. Pairwise comparison is one of the most common ways to learn the order or rank relationship among samples. For instance, [20] employed a
Figure 1. The overview of the proposed Order-ViT during training. For the optimization of Order-ViT, we utilize a feature extractor, a categorical classifier, and an order classifier. The categorical classifier conducts classification using the input image alone. The order classifier compares the input image against the reference image. pariwise comparator to compare two samples and to determine their ordering relationship. [36] developed RankNet that learns a ranking function to predict the ranking posterior of two samples. These methods generally utilize the absolute rank of a sample. In order to learn the relative rank that quantifies the degree of difference, [28] proposed a regression-based order learning mechanism for age estimation. Although it has been successfully applied to age estimation, it has not been utilized for pathology image analysis. To the best of our knowledge, this is the first attempt to incorporate order learning into cancer diagnosis in pathology images. Ordinal classification or regression approaches, which aim at directly predicting the class label of a sample on an ordinal scale, have been adopted for pathology image analysis; for example, [18] conducted both categorical classification and regression for prostate and colorectal cancer grading. However, the ordering relationships among samples and class labels have not been explicitly utilized for cancer diagnosis in pathology images.

3. Methods
3.1. Problem Formulation
Suppose that we are given \( N \) pathology images and their ground truth labels \( \{x_i, y_i\}_{i=1}^{N} \), where \( x_i \in \mathbb{R}^3 \) is the \( i \)-th pathology image, \( y_i \in C = \{c_1, c_2, \ldots, c_N_c\} \) is the ground truth class label for \( x_i \). \( c_1, c_2, \ldots, c_{N_c} \) represent cancer grades, of which each owns a unique histopathology pattern, and are also defined on an ordinal scale \( c_1 = 1 < c_2 = 2 < \ldots < c_{N_c} = N_c \). Given \( \{x_i, y_i\}_{i=1}^{N} \), Order-ViT maps the images into a high-dimensional embedding space in which we conduct two classification tasks, i.e., categorical classification and order classification. The objective of our study is to learn Order-ViT that can leverage the ordering relationships among pathology images and conduct cancer grading in an accurate and robust manner, which can be formulated as follows:

\[
\min \sum_{i=1}^{N} L(\xi(x_i), y_i; \theta)
\]

where \( L \) is the loss function and \( \theta \) is a set of learnable parameters of Order-ViT \( \xi \).

3.2. Order-ViT
Order-ViT \( \xi \) contains four major computational blocks: 1) a feature extractor \( \xi^{feat} \), a categorical classifier \( \xi^{cat} \), an order classifier \( \xi^{ord} \), and a voting block \( \xi^{vot} \). Given an input image \( x \), the feature vector \( \xi^{feat} \) extracts a feature vector \( f \), which is subsequently fed into the categorical classifier \( \xi^{cat} \) for conducting categorical classification, i.e., directly predicting the class label for \( x \). For order learning, a reference image \( x^{ref} \) goes through \( \xi^{feat} \), generating a reference feature vector \( f^{ref} \). \( f \) and \( f^{ref} \) are concatenated together and fed into the order classifier \( \xi^{ord} \) to predict the order relationship between \( x \) and \( x^{ref} \). The voting block \( \xi^{vot} \) is invoked at inference only to identify hard/uncertain images and to correct and stabilize the prediction.

3.2.1 Feature Extractor
The feature extractor \( \xi^{feat} \) is built based upon a vision transformer (ViT). Suppose that we are given an input image \( x \), \( \xi^{feat} \) adopts Resnet50V2 [9] to extract the initial feature vector \( e \in \mathbb{R}^{W \times H \times N_e} \) (\( W = H = 24 \), \( N_e = 1024 \)) and generates a set of token embeddings via a linear layer \( \{s_i|i = 1, 2, \ldots, N_s\}, s_i \in \mathbb{R}^{N_f} \) where \( N_s \) is the length of the token embeddings (\( N_s = 576 \)) and \( N_f \) is the size of each token embedding (\( N_f = 768 \)). For classification,
we incorporate a class token $s^{CLS} \in \mathcal{R}^N$ into the token embeddings, producing $H_0 = [s_0, s_1, s_2, ..., s_N], s_0 = s^{CLS}$. We also utilize positional embedding to add location information for each token embedding in a way that $\sin(i/10000^{2j+1/N})$ and $\cos(i/10000^{2j+1/N})$ are added to every 2j-th and 2j + 1-th dimensions of $s_i, i = 0, 1, ..., N$. The resultant token embeddings are fed to the Transformer encoder, which is built based upon multi-head self-attention (MHSA) layers, feed-forward neural network (\textit{FFNN}(\cdot)) layers, and a layer normalization (LN) layer, to extract the output feature $h_{l}$ as follows:

$$h'_{l} = \text{MHSA}(\text{LN}(h_{l-1})) + h_{l-1}, l = 1, ..., N_l$$

$$h_{l} = \text{FFNN}(\text{LN}(h_{l})) + h'_{l}, l = 1, ..., N_l$$

$$f = \text{LN}(h^0_N)$$

where $B_l$ is the number of Transformer encoders ($N_l = 12$) and $h^0_N$ is the output embedding corresponding to $s^{CLS}$ at the $N_l$-th Transformer encoder. MHSA is composed of multiple, parallel self-attentions (SAs) as follows:

$$\text{MHSA}(h) = \text{Linear}([\text{SA}_1(h), ..., \text{SA}_m(h)])$$

$$\text{SA}_i(h) = \text{softmax}(qk^T/\sqrt{D})v, i = 1, ..., N_m$$

where $q, k, v = \text{Linear}(h) \in \mathcal{R}^{d \times m}$ are query, key, value vectors, respectively. Linear is a linear layer, $N_m$ is the number of SAs, and $D = N_l/N_m$. FFNN contains $\text{Linear} - \text{ReLU} - \text{Linear} - \text{ReLU} - \text{Linear}$.

### 3.2.2 Categorical Classifier

The categorical classifier $\xi^{cat}$ contains $\text{Linear} - \text{LeakyReLU} - \text{Linear} - \text{LeakyReLU} - \text{Linear}$. Provided with a feature vector $f$, $\xi^{cat}$ produces the posterior probabilities for the class labels $p = \{\hat{p}_c | c = 1, ..., C_c\}$ where $\hat{p}_c$ denotes the probability that the input image belongs to the $c$-th class label.

### 3.2.3 Order Classifier

Following [20], the order relationship between two samples $x_i$ and $x_j$ can be determined by using their class labels $y_i$ and $y_j$ as follows:

$$x_i < x_j \quad \text{if} \quad y_i - y_j < 0$$

$$x_i \approx x_j \quad \text{if} \quad y_i - y_j = 0$$

$$x_i > x_j \quad \text{if} \quad y_i - y_j > 0$$

where $x_i \succ x_j, x_i \approx x_j,$ and $x_i < x_j$ indicate that $x_i$ is greater than, equal to, or less than $x_j$ with respect to the class labels that are defined on an ordinal scale. In the context of cancer grading, these are interpreted as the grade of $x_i$ is higher than, equal to, or lower than that of $x_j$. Given the two feature vectors from $x_i$ and $x_j$, the objective of the order classifier $\xi^{ord}$ is to predict the order relationship ($\succ, \approx, \prec$) between $x_i$ and $x_j$ that is defined as follows:

$$z_{i,j} = \begin{cases} r_1, & \text{if} \ y_i < y_j \\ r_2, & \text{if} \ y_i = y_j \\ r_3, & \text{if} \ y_i > y_j \end{cases}$$

where $r_1, r_2,$ and $r_3 \in \mathcal{N}$ are the ground truth order class labels that correspond to $\succ, \approx,$ and $\prec$, respectively.

During training, each image in a batch is designated as an anchor image $x^a$ and $\xi^{cat}$ is used to generate an anchor feature vector $f^a$. Then, a reference feature vector $f^{ref}$ is selected from a feature memory bank $M$ and is concatenated with $f^a$, producing the order feature vector $f^{ord} = f^a \oplus f^{ref}$. $x^{ord}_i$ receives $f^{ord}$ and produces the posterior probabilities for the order class labels $\hat{o} = \{\hat{o}_i | i = 1, ..., N_0\}$ where $\hat{o}_i$ denotes the probability that the input image belongs to the $i$-th order class label.

The architecture of $\xi^{ord}$ is identical to that of $\xi^{cat}$ except the number of neurons in the last linear layer.
3.3. Two-stage Inference

During inference, we make the final decision or prediction in two stages. In the first stage, we make the prediction per input image $x^a$ using the categorical classifier $\xi^{cat}$. Specifically, we extract the feature vector $f^a$ for $x^a$ by using the feature extractor $\xi^{feat}$. Then, $f^a$ is fed into $\xi^{cat}$ to conduct categorical classification. Using the output of $\xi^{cat}$, we also calculate a confidence score $u$ of the input image. If the prediction is certain, i.e., the confidence score $u > \tau$, then the prediction by $\xi^{cat}$ becomes the final prediction. Otherwise, we invoke a voting block $\xi^{vot}$ in the second stage. $\xi^{vot}$ utilizes the order classifier $\xi^{ord}$ using a feature memory bank $\mathcal{M} = \{\mathcal{M}^c | c = c_1, c_2, ..., c_{N_c}\}$ where $\mathcal{M}^c$ is a set of feature vectors of class $c$. For each feature vector $f^m \in \mathcal{M}^c$, we create the order feature vector $f^{ord} = f^a \oplus f^m$ and fed it into the order classifier $\xi^{ord}$, producing the order relationship ($\succ$, $\approx$, $\prec$) between them. We accumulate the order classification results for $\mathcal{M}$ and determine the final prediction.

3.3.1 Confidence Score

We compute the confidence score $u$ of the input image by obtaining the posterior probabilities for the class labels $\hat{p} = \{\hat{p}_i | i = 1, ..., N_c\}$, given by $\xi^{cat}$, identifying the highest probability $\hat{p}_{first}$ and the second highest probability $\hat{p}_{second}$, and calculating the difference between them, which is designated as the confidence score $u = \hat{p}_{first} - \hat{p}_{second}$. The larger the difference between the two probabilities is, the more confident the decision is.

To determine whether we invoke the voting scheme $\xi^{vot}$ or not, we calculate the threshold value $\tau$ using the images in the validation set. From the validation set, we collect the images that are mis-classified by $\xi^{cat}$, which forms $\{x_i | i = 1, ..., N_m\}$ where $N_m$ is the number of mis-classified images in the validation set. For these images, we compute the confidence scores $\{u_i | i = 1, ..., N_m\}$ as described above and sort the images in decreasing order by the confidence scores. Using each confidence score as a threshold value $\tau^*$, we obtain the images that are smaller than $\tau^*$, i.e., the images that the model is less confident. The number of such images is designated as $N_m$. Then, we select the threshold value $\tau$ as the largest $\tau^*$ that produces $\frac{N_m}{N_m} \leq 0.1$.

3.3.2 Feature Memory Bank

The feature memory bank $\mathcal{M} = \{\mathcal{M}^c | c = c_1, c_2, ..., c_{N_c}\}$ consists of a number of feature vectors per class label. To construct $\mathcal{M}$, we select $N_r$ reference images per class label from the training and validation sets that are representative of each class. $\xi^{cat}$ classifies each image in the training and validation sets. If the result is correct and its highest probability is 0.9 or higher, then it is designated as a candidate sample. Once we obtain all the candidate samples, we randomly select $N_r$ images per class label, extract their feature vectors, and construct $\mathcal{M}$. We set $N_r$ to 10.

3.3.3 Voting

Given $f^a$ of an uncertain image, i.e., anchor image $x^a$, and $\mathcal{M}$, $\xi^{vot}$ uses $\xi^{ord}$ to conduct a pair-wise comparison between $x^a$ and every reference image $x^{ref}_c$ that is represented by the feature vector in $\mathcal{M}$. The results of $\xi^{ord}$, i.e., $\succ$, $\approx$, and $\prec$ indicate that $x^a$ is ranked higher than, equal to, and lower than $x^{ref}_c$, respectively. Suppose that the ground truth class label of $x^{ref}_c$ is $c$. We vote for the prediction in three different manners: 1) if $\xi^{ord}$ produces $\succ$, we add $1/(N_c - c)$ for all labels higher than $c$, 2) if the result is $\approx$, we add $1/(N_c - c)$ for all labels lower than $c$, and 3) if the result is $\prec$, we add 1.0 to $c$. By normalizing the votes, we construct the $q = \{q^c | c = 1, ..., N_c\}$ where $q^c$ represents the probability that $x^a$ belongs to the $c$-th class label.

For $x^a$, two posterior probabilities are available, including $\hat{p}$ by $\xi^{cat}$ and $q$ by $\xi^{vot}$. To make the final prediction, we conduct a weighted sum of the two posterior probabili-
ties and take the class label with the highest probability. The weights ($w_1$ and $w_2$) are adaptively computed as follows:

$$w_1 = 1 - \frac{H(\hat{p})}{H(q) + H(\hat{p})}, \quad w_2 = 1 - \frac{H(q)}{H(q) + H(\hat{p})}$$

(9)

where $H$ is an entropy. The less uncertain the probability distribution is, the higher weight it gets. The final probabilities $q_{\text{final}}$ are calculated as $q_{\text{final}} = \hat{p} \times w_1 + q \times w_2$.

3.4. Loss Function

The objective function for Order-ViT is given by

$$L_{\text{total}} = L_{\text{cat}} + \lambda L_{\text{ord}}$$

(10)

where $L_{\text{cat}}$ and $L_{\text{ord}}$ denotes the loss function for category classification and order classification, respectively, and $\lambda$ is a hyper-parameter to balance the effect of the two loss functions ($\lambda = 0.4$). Both category classification and order classification adopt cross-entropy loss.

4. Experiment

Table 1. Details of colorectal and gastric tissue datasets.

<table>
<thead>
<tr>
<th>Tissue Type</th>
<th>Class</th>
<th>Training</th>
<th>Validation</th>
<th>Test I</th>
<th>Test II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorectal Tissue</td>
<td>BN</td>
<td>773</td>
<td>374</td>
<td>453</td>
<td>27,896</td>
</tr>
<tr>
<td></td>
<td>WD</td>
<td>1,866</td>
<td>264</td>
<td>192</td>
<td>8,394</td>
</tr>
<tr>
<td></td>
<td>MD</td>
<td>2,997</td>
<td>370</td>
<td>738</td>
<td>61,985</td>
</tr>
<tr>
<td></td>
<td>PD</td>
<td>1,391</td>
<td>234</td>
<td>205</td>
<td>11,896</td>
</tr>
<tr>
<td>Gastric Tissue</td>
<td>BN</td>
<td>20,883</td>
<td>8,398</td>
<td>7,955</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WD</td>
<td>14,251</td>
<td>2,239</td>
<td>1,795</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD</td>
<td>20,815</td>
<td>2,370</td>
<td>2,458</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PD</td>
<td>27,689</td>
<td>2,374</td>
<td>3,579</td>
<td></td>
</tr>
</tbody>
</table>

4.1. Dataset

We employ two kinds of pathology image datasets, including colorectal tissue dataset and gastric tissue dataset. The details of the datasets are described in Table 1.

4.2. Colorectal Tissue Dataset

Two sets of colorectal tissue datasets were obtained from [18]. The first data set was obtained from 6 colorectal tissue microarrays (TMAs) (0.2465 μm × 0.2465 μm) and 3 WSIs (0.2518 μm × 0.2518 μm), at 40x magnification using an Aperio digital slide scanner (Leica Biosystems). This contains a training set, a validation set, and a test set ($C_{\text{TestI}}$). The second dataset ($C_{\text{TestII}}$) was obtained from 45 WSIs (0.2253 μm × 0.2253 μm) at 40 × magnification using a NanoZoomer digital slide scanner (Hamamatsu Photonics K.K.). For both datasets, tissue image patches (1024 × 1024 pixels) are annotated with benign (BN), well-differentiated (WD), moderately differentiated (MD), and poorly-differentiated (PD).

4.3. Gastric Tissue Dataset

98 anonymized, gastric WSIs, digitized at 40 × magnification using an Aperio digital slide scanner (Leica Biosystems) (0.2635 μm × 0.2635 μm), were obtained from a local hospital. Upon pathologic review, 114,842 image patches (1024 × 1024 pixels) were obtained and categorized into BN, tubular well-differentiated (TW) tumor, tubular moderately-differentiated (TM), and tubular poorly-differentiated (TP) tumor. These constitute a training set, a validation set, and a test set ($G_{\text{Test}}$).

4.4. Comparative Experiments

We compare Order-ViT with a number of CNN- and Transformer-based models. Three types of CNN-based models are employed: 1) three plain CNNs: ResNet50 [9], DenseNet121 [12], and EfficientNetB0 [31], 2) a multi-scale CNN: MSBP-Net [33], and 3) two multi-task CNNs [18]: $M_{\text{MSE-CE}}$ and $M_{\text{MAE-CE}}$. Three Transformer-based models are utilized: 1) ViT [7], 2) Swin-transformer [22], and 3) DeiT III [32].

4.5. Evaluation Metrics

Three evaluation metrics are adopted: 1) accuracy (Acc): measures the ratio of correctly classified instances to the total number of instances, 2) macro F1-score (F1): a harmonic mean of the average precision and average recall, and 3) quadratic weighted kappa ($k_{\text{qw}}$): quantifies the degree of agreement between the prediction and ground truth labels.

4.6. Implementation Details

We implemented all models using the Pytorch platform with two RTX 3090 GPUs. All the models adopted pre-trained weights from ImageNet and were trained with 50 epochs using a batch size of 16. Adam optimizer ($\beta_1 : 0.9, \beta_2 : 0.999, e : 1.0e^{-8}$), and Cosine annealing warm restarts scheduler (learning rate: $1.0e^{-4}$). All the tissue images were resized to 384 x 384 pixels and 512 x 512 pixels Transformer-based and CNN-based models, respectively. Several data augmentation techniques, including Affine transformation, Gaussian blur, average blur, median blur, Gaussian noise, dropout, linear contrast, horizontally and vertically flipping, and random saturation, were applied using Aleju library.

5. Result

5.1. Colorectal Cancer Classification

Table 2 shows the results of colorectal cancer classification ($C_{\text{TestI}}$ and $C_{\text{TestII}}$) by Order-ViT and other competing models. On $C_{\text{TestI}}$, Order-ViT achieved the best Acc of 87.66% and $k_{\text{qw}}$ of 0.942 and the second best F1 of 0.834, outperforming all competitors except F1 by
\( M_{\text{MSE-CE}_w} \). Among the competitors, two multitask CNN models (\( M_{\text{MSE-CE}_w} \) and \( M_{\text{MAE-CE}_w} \)) were superior to others. Among Transformer-based models, ViT outperformed two other Transformer-based models. DeiT III was inferior to all models regardless of the evaluation metrics. On \( C_{\text{Test} I} \), Order-ViT outperformed all competitors regardless of the evaluation metrics. Among other competitors, two multitask CNN models (\( M_{\text{MSE-CE}_w} \) and \( M_{\text{MAE-CE}_w} \)) and ViT were still the best performing models. We note that \( C_{\text{Test} II} \) was obtained under different acquisition settings from the training set. In a head-to-head comparison between the two datasets, there was a consistent performance drop for all models under consideration. However, Order-ViT showed the least performance drop in comparison to other models. For instance, 4.45% Acc, 0.094 F1, and 0.043 \( k_w \) were dropped by Order-ViT. 8.16 \( \sim \) 28.34% Acc, 0.099 \( \sim \) 0.277 F1, and 0.057 \( \sim \) 0.523 \( k_w \) were dropped by other models.

### 5.2. Gastric Cancer Classification

Table 2 describes the performance of Order-ViT and other competitors on gastric cancer classification (\( G_{\text{Test}} \)). Order-ViT achieved the best Acc of 84.89% and F1 of 0.783 and second best \( k_w \) of 0.930. Among other models, mixed results were observed; MSBP-Net, ResNet, and ViT obtained the best Acc of 84.04%, F1 of 0.775, and \( k_w \) of 0.931, respectively. DeiT III was still the worst model in all evaluation metrics.

### 5.3. Ablation Study

We conducted a series of ablation experiments on Order-ViT to investigate the effect and role of its computational components and design choices.

**Effect of computational blocks.** First, we assessed the effect of the computational blocks, including the order classifier \( \xi^{\text{ord}} \) and/or the voting block \( \xi^{\text{vot}} \). on Order-ViT. Table 3 demonstrates the results of colorectal and gastric cancer classification without \( \xi^{\text{ord}} \) and/or \( \xi^{\text{vot}} \). Order-ViT without \( \xi^{\text{vot}} \) experienced a consistent performance drop across different datasets; 0.25% Acc, 0.005 F1, and 0.001 \( k_w \) on \( C_{\text{Test} I} \); 0.44% Acc, 0.003 F1, and 0.002 \( k_w \) on \( C_{\text{Test} II} \); 0.31% Acc, 0.007 F1, and 0.001 \( k_w \) on \( G_{\text{Test}} \). Eliminating both \( \xi^{\text{ord}} \) and \( \xi^{\text{vot}} \), the performance, in general, further decreased except \( k_w \) on \( G_{\text{Test}} \); 1.39% Acc, 0.005 F1, and 0.011 \( k_w \) on \( C_{\text{Test} II} \), 5.67% Acc, 0.028 F1, and 0.025 \( k_w \) on \( C_{\text{Test} II} \), and 0.83% Acc and 0.011 F1 on \( G_{\text{Test}} \). Moreover, we assessed the effect of the weighted sum of the two posterior probabilities from the categorical and order classifications (\( \Sigma_w \)). Replacing \( \Sigma_w \) by addition, there is a slight performance drop regardless of the datasets and evaluation metrics. These results demonstrate that the effectiveness of the computational blocks in Order-ViT.

**Effect of the number of reference images.** Second, we investigated the effect of the number of reference images \( N_r \) in the voting block \( \xi^{\text{vot}} \). We set \( N_r \) to 10, 30, 50, 70, 90, 100, and 200. As \( N_r \) changes, the classification performance was more or less the same across different datasets. This suggests that Order-ViT is not sensitive to the number of references images for voting. By setting \( N_r \) to 10, we can not only improve the accuracy and robustness of cancer classification, but also reduce the computational complexity during the inference.

### 5.4. Computational Complexity

Table 5 depicts the computational complexity of Order-ViT and other competing models. Including Order-ViT, Transformer-based models contain 2- to 18-fold more parameters than CNN-based models. Similar observations were made for the number of floating-point operations (FLOPs). The training time of the models varied regardless of the type of the models; however, CNN-based models, in general, demonstrated the shorter inference time than Transformer-based models. Due to the hybrid architecture of the feature extractor and voting mechanism, Order-ViT showed the most FLOPs and the longest inference time among the models under considerations.

![Table 2. Result of colorectal and gastric cancer classification.](image-url)

<table>
<thead>
<tr>
<th>Model</th>
<th>( C_{\text{Test} I} )</th>
<th>( C_{\text{Test} II} )</th>
<th>( G_{\text{Test}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc(%)</td>
<td>F1</td>
<td>( k_w )</td>
</tr>
<tr>
<td>ResNet</td>
<td>86.65</td>
<td>0.825</td>
<td>0.935</td>
</tr>
<tr>
<td>DenseNet</td>
<td>85.77</td>
<td>0.820</td>
<td>0.930</td>
</tr>
<tr>
<td>EfficientNet</td>
<td>86.71</td>
<td>0.821</td>
<td>0.926</td>
</tr>
<tr>
<td>MSBP-Net</td>
<td>86.21</td>
<td>0.824</td>
<td>0.933</td>
</tr>
<tr>
<td>( M_{\text{MSE-CE}_w} )</td>
<td>87.28</td>
<td>0.838</td>
<td>0.940</td>
</tr>
<tr>
<td>( M_{\text{MAE-CE}_w} )</td>
<td>86.65</td>
<td>0.826</td>
<td>0.937</td>
</tr>
<tr>
<td>ViT</td>
<td>86.27</td>
<td>0.829</td>
<td>0.931</td>
</tr>
<tr>
<td>Swin</td>
<td>85.26</td>
<td>0.820</td>
<td>0.931</td>
</tr>
<tr>
<td>DeiT III</td>
<td>76.76</td>
<td>0.673</td>
<td>0.794</td>
</tr>
<tr>
<td>Order-ViT (Ours)</td>
<td><strong>87.66</strong></td>
<td>0.834</td>
<td><strong>0.942</strong></td>
</tr>
</tbody>
</table>
Table 3. Ablation study on Order-ViT for colorectal and gastric cancer classification.

<table>
<thead>
<tr>
<th>$\xi_{cat}$</th>
<th>$\xi_{ord}$</th>
<th>$\xi_{vot}$</th>
<th>$\sum w$</th>
<th>$C_{Test I}$</th>
<th>$C_{Test II}$</th>
<th>$G_{Test}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>Acc(%) F1 k_w</td>
<td>Acc(%) F1 k_w</td>
<td>Acc(%) F1 k_w</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>86.27</td>
<td>0.829</td>
<td>0.931</td>
<td>77.54</td>
<td>0.712</td>
<td>0.874 84.06 0.772 0.931</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>87.41</td>
<td>0.829</td>
<td>0.941</td>
<td>82.77</td>
<td>0.737</td>
<td>0.897 84.58 0.776 0.929</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓</td>
<td>87.47</td>
<td>0.831</td>
<td>0.941</td>
<td>82.98</td>
<td>0.739</td>
<td>0.898 84.61 0.779 0.929</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>87.66</td>
<td>0.834</td>
<td>0.942</td>
<td>83.21</td>
<td>0.740</td>
<td>0.899 84.89 0.783 0.930</td>
</tr>
</tbody>
</table>

Table 4. Effect of $N_r$ on colorectal and gastric cancer classification.

<table>
<thead>
<tr>
<th>$N_r$</th>
<th>$C_{Test I}$</th>
<th>$C_{Test II}$</th>
<th>$G_{Test}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>87.66 0.834 0.942</td>
<td>83.21 0.740 0.899</td>
<td>84.89 0.783 0.930</td>
</tr>
<tr>
<td>30</td>
<td>87.47 0.831 0.940</td>
<td>83.20 0.740 0.899</td>
<td>84.87 0.782 0.929</td>
</tr>
<tr>
<td>50</td>
<td>87.47 0.831 0.940</td>
<td>83.20 0.740 0.899</td>
<td>84.85 0.782 0.929</td>
</tr>
<tr>
<td>70</td>
<td>87.59 0.832 0.941</td>
<td>83.20 0.740 0.899</td>
<td>84.88 0.782 0.929</td>
</tr>
<tr>
<td>90</td>
<td>87.59 0.832 0.941</td>
<td>83.20 0.740 0.899</td>
<td>84.85 0.782 0.929</td>
</tr>
<tr>
<td>100</td>
<td>87.53 0.832 0.941</td>
<td>83.19 0.739 0.899</td>
<td>84.88 0.782 0.929</td>
</tr>
<tr>
<td>200</td>
<td>87.53 0.832 0.941</td>
<td>83.19 0.739 0.899</td>
<td>84.88 0.782 0.929</td>
</tr>
</tbody>
</table>

Table 5. Computational complexity of Order-ViT and other competing models.

<table>
<thead>
<tr>
<th>Model</th>
<th>FLOPs (M)</th>
<th>Params (M)</th>
<th>Training (ms/image)</th>
<th>Inference (ms/image)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet</td>
<td>21587.17</td>
<td>24.69</td>
<td>18.13</td>
<td>3.09</td>
</tr>
<tr>
<td>DenseNet</td>
<td>15954.34</td>
<td>7.35</td>
<td>5.71</td>
<td>4.42</td>
</tr>
<tr>
<td>EfficientNet</td>
<td>152.22</td>
<td>4.81</td>
<td>3.31</td>
<td>3.12</td>
</tr>
<tr>
<td>MSBP-Net</td>
<td>44285.04</td>
<td>29.50</td>
<td>47.37</td>
<td>9.21</td>
</tr>
<tr>
<td>$M_{MSE-CE}$</td>
<td>141.14</td>
<td>4.84</td>
<td>27.44</td>
<td>4.28</td>
</tr>
<tr>
<td>$M_{MAE-CE}$</td>
<td>141.14</td>
<td>4.84</td>
<td>27.77</td>
<td>4.19</td>
</tr>
<tr>
<td>ViT</td>
<td>49505.12</td>
<td>86.37</td>
<td>27.01</td>
<td>7.84</td>
</tr>
<tr>
<td>Swin</td>
<td>44753.49</td>
<td>87.33</td>
<td>45.51</td>
<td>8.45</td>
</tr>
<tr>
<td>DeiT III</td>
<td>49608.92</td>
<td>86.17</td>
<td>25.84</td>
<td>6.27</td>
</tr>
<tr>
<td>Order-ViT</td>
<td>99046.44</td>
<td>87.29</td>
<td>27.95</td>
<td>9.76</td>
</tr>
</tbody>
</table>

5.5. Conclusions

In this work, we propose an efficient and effective cancer classification method that can leverage the individual histological patterns of pathology images by categorical classification and the relationship among different pathology images by order learning. The experimental results demonstrate that the proposed method can aid in improving the accuracy and robustness of cancer classification in pathology images. The future study will entail further development of order learning approaches for pathology image analysis and application of the proposed method to other types of organs and diseases.

Acknowledgements

This work was supported by the grant of the National Research Foundation of Korea (NRF) (No. 2021R1A2C2014557) and Institute of Information & communications Technology Planning & evaluation (IITP) (No. RS-2022-00167143), funded by the Korea government (MSIT).

References

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,


[18] Trinh Thi Le Vuong, Kyungeun Kim, Boram Song, and Jin Tae Kwak. Joint categorical and ordinal learning for cancer grading in pathology images. Medical image analysis, 73:102206, 2021. 1, 3, 6


