Towards a text-based quantitative and explainable histopathology image analysis

Anh Tien Nguyen, Trinh Thi Le Vuong, and Jin Tae Kwak

School of Electrical Engineering, Korea University, Seoul 02841, Korea
{ngtienanh,trinhvg,jkwak}@korea.ac.kr

Abstract. Recently, vision-language pre-trained models have emerged in computational pathology. Previous works generally focused on the alignment of image-text pairs via the contrastive pre-training paradigm. Such pre-trained models have been applied to pathology image classification in zero-shot learning or transfer learning fashion. Herein, we hypothesize that the pre-trained vision-language models can be utilized for quantitative histopathology image analysis through a simple image-to-text retrieval. To this end, we propose a Text-based Quantitative and Explainable histopathology image analysis, which we call TQx. Given a set of histopathology images, we adopt a pre-trained vision-language model to retrieve a word-of-interest pool. The retrieved words are then used to quantify the histopathology images and generate understandable feature embeddings due to the direct mapping to the text description. To evaluate the proposed method, the text-based embeddings of four histopathology image datasets are utilized to perform clustering and classification tasks. The results demonstrate that TQx is able to quantify and analyze histopathology images that are comparable to the prevalent visual models in computational pathology. The repository is available at https://github.com/QuIIL/TQx.

Keywords: Computational pathology · Vision-language model · Image-to-text retrieval.

1 Introduction

Over the past years, the success of deep learning models has been attributable to the single-modal pre-trained models in computer vision and natural language processing. In computer vision, various convolutional neural networks (CNNs) [1–3] and Vision Transformer (ViT) [4–6], known as vision models, are pre-trained on ImageNet and have been successfully applied to a wide range of downstream tasks such as image classification, retrieval, and segmentation [7]. In natural language processing, several pre-trained language models such as BERT [8] and GPT [9] are available and have shown to be effective in many text-based applications, including text classification, question answering, translation, and summarization [10]. Recently, there exists an emerging practice of fusing the vision and natural language models, which forms vision-language models

(VLMs). VLMs are, in general, pre-trained on an extensive collection of image-text datasets by jointly learning visual and textual modalities/knowledge. A notable example is CLIP [11], which simultaneously optimizes vision and language models to align the visual and textual representations. Computational pathology is not an exception. PLIP [12] and QUILT-Net [13] are two exemplary VLMs that were pre-trained on massive histopathology datasets with pairs of histopathology image and histopathology description via the contrastive learning paradigm similar to CLIP. Such VLMs build a solid foundation for various types of downstream tasks. In particular, these models have been successfully applied to several image classification tasks such as lymph-node metastasis detection, tissue phenotyping, and Gleason grading [12, 13] without further training or fine-tuning, i.e., zero-shot image classification.

We have observed that the existing works on VLMs have mainly focused on the pre-training process and straightforward, direct application to downstream tasks [11–14]. Some have sought to utilize VLMs with transfer learning and knowledge distillation [15, 16]. These approaches are, by and large, similar to the way the pre-trained vision models and language models are used, which do not fully explore the potential of VLMs. Herein, we hypothesize that the pre-trained VLMs *per se* are capable of conducting quantitative histopathology image analysis. In other words, VLMs are able to quantify histopathology images, and the resultant quantitative features can be used for downstream tasks. To test our hypothesis, we propose a Text-based Quantitative and Explainable histopathology image analysis framework, called TQx. We systematically evaluate the effectiveness of TQx using four histopathology image datasets by performing clustering and classification tasks. The experimental results suggest that TQx not only provides a capacity for quantifying and analyzing histopathology images comparable to the conventional vision models but also permits the direct interpretation of the results with human-readable words.

2 Methodology

TQx involves two major components: 1) a pre-trained VLM and 2) a word-of-interest (WoI) pool. The VLM has a text encoder and a visual encoder that was jointly optimized via contrastive learning [11]. The WoI pool includes a set of pathology terms that explain the characteristics of histopathology images. Overall, by utilizing cosine similarity, the VLM retrieves relevant keywords from the WoI pool, then the text-based image embedding is generated from these keywords (Fig. 1). The most important step is to filter the related terms because they directly affect the generation of the text-based representation. Therefore, in the following sections, we will explain this procedure in detail.

2.1 Text-based image representation

Suppose that we are given a pre-trained VLM $\mathcal{V}$, a set of pathology images $\mathcal{X} = \{x_i\}_{i=1}^{N}$, and a WoI pool $\mathcal{W} = \{w_j\}_{j=1}^{N_w}$ where x_i is the ith histopathology

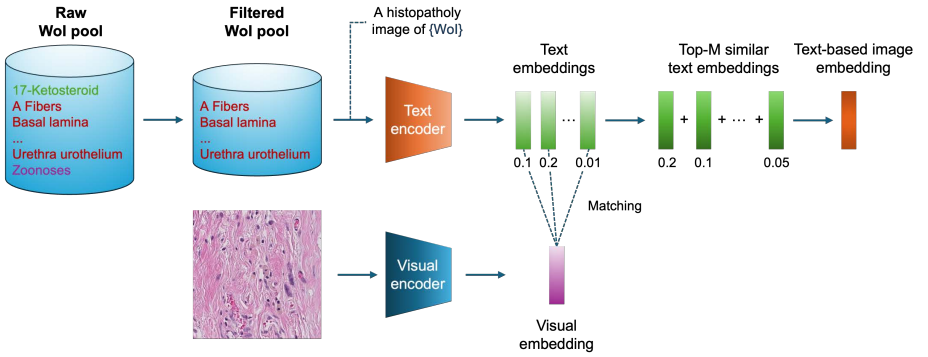

Fig. 1. The raw WoI pool stores all UMLS [17] pathology terms of various semantic types. The filtered pool is obtained by selecting a particular semantic type under consideration. The pair of encoders from a pre-trained VLM generates text and visual embeddings, which are then compared together. The similarity scores from the comparison are normalized and then used as weights to produce a text-based embedding.

image, w_j is the jth keyword, and N and N_w are the number of histopathology images and keywords, respectively. Each x_i undergoes the visual encoder of $\mathcal{V}$ to produce the visual embedding $\mathbf{v}_i$, and each w_j is fed into the text encoder of $\mathcal{V}$ to generate the corresponding text embedding $\mathbf{f}_j$. For all $(\mathbf{v}_i, \mathbf{f}_j)$, we compute the similarity scores, producing $\mathcal{S} = \{s_{i,j} | i = 1, ..., N \wedge j = 1, ..., N_w\}$ where $s_{i,j}$ denotes the similarity between x_i and w_j. Using $\mathcal{S}$, we select the top-M keywords $\mathcal{W}^M = \{w'_i\}_{i=1}^M$ that are most representative of $\mathcal{X}$. For each image x_i, we first compute the rank of the keywords in $\mathcal{W}$ as follows: $\forall(j,k)$, $r_{i,j} < r_{i,k}$ if $s_{i,j} < s_{i,k}$ where $r_{i,j} \in \{1, ..., N_w\}$ is the rank of w_j for x_i. Then, we average the rank of the keywords among $\mathcal{X}$ to select a set of M keywords with the highest ranks, designated as $\mathcal{W}^M = \{w'_i\}_{i=1}^M$, and use it to produce the corresponding text embeddings $\mathcal{F}^M = \{\mathbf{f}'_j\}_{j=1}^M$ and their similarity scores $\mathcal{S}^M = \{s'_{i,j} | i = 1, ..., N \wedge j = 1, ..., M\}$.

To obtain the text-based image representation for each x_i, we normalize the similarity scores between x_i and all the keywords in $\mathcal{W}^M$, i.e., $\{s'_{i,j}\}_{j=1}^M$, using the softmax operation to produce weights for the text embeddings $\{\alpha_j\}_{j=1}^M$. Then, we compute a weighted sum of the text embeddings to generate the final text-based image embedding of x_i given by $\mathbf{f}_i^T = \sum_{j=1}^M \alpha_j \mathbf{f}'_j$. The text-based image embedding $\mathbf{f}^T$ delivers the abstract textual information of the input image, which can be used for quantitative analysis. The top-M keywords $\mathcal{W}^M$ are human-readable and -understandable, thus enabling the interpretation of the results. Therefore, the text-based image embedding is self-explainable *per se*.

2.2 Construction of word-of-interest pool

To construct the WoI pool, we utilize QUILT-1M [13], the largest image-text dataset in pathology containing a wide range of pathology keywords. First, we

collect all keywords (called entities) from QUILT-1M, then cross-check and process with standard terms in the Unified Medical Language System (UMLS) [17]. After eliminating the duplicate terms using the UMLS concept unique identifier (CUI), we obtain the raw WoI pool of 28,292 keywords, which is designated as $\mathcal{W}_{Level-0}$.

2.3 Refinement of word-of-interest pool

In the raw WoI pool $\mathcal{W}_{Level-0}$, each keyword is associated with the corresponding UMLS semantic type that defines its category, such as *Tissue*, *Cell*, or *Neoplastic Process*. In the UMLS, these entities are categorized based on the tree-like semantic network, and thus, one semantic type may include one or more sub-types or -groups. For example, *Fully Formed Anatomical Structure* contains multiple smaller sub-groups such as *Cell*, *Tissue*, and *Cell Component*. Each semantic (sub-)type has its own histopathology meaning.

To investigate the impact of the WoI pool and the semantic category on TQx, we choose three UMLS semantic types that are relevant to pathology diagnosis: 1) *Neoplastic Process*: abnormal growth of tissue, 2) *Disease or Syndrome*: an abnormal condition of an organism, and 3) *Pathologic Function*: a disordered process, activity, or state of the organism or part of it. Then, we construct three WoI pools including $\mathcal{W}_{Level-3}$ (*Neoplastic Process*), $\mathcal{W}_{Level-2}$ (*Disease or Syndrome*), and $\mathcal{W}_{Level-1}$ (*Pathologic Function*) with 2,215, 5,441, and 6,232 keywords, respectively. In total, we build four WoI pools. For each of the four WoI pools, we retrieve the top-M highly ranked keywords with $M = 1000$, resulting in $\mathcal{W}^M_{Level-0}$, $\mathcal{W}^M_{Level-1}$, $\mathcal{W}^M_{Level-2}$, and $\mathcal{W}^M_{Level-3}$. $\mathcal{W}^M_{Level-0}$ (raw WoI pool) is the most inclusive and general while $\mathcal{W}^M_{Level-3}$ (*Neoplastic Process*) is the most confined and specific. Moreover, using the four WoI pools, we generate the four sets of text-based feature embeddings: $\mathcal{F}^{\mathcal{W}^M_{Level-0}}$, $\mathcal{F}^{\mathcal{W}^M_{Level-1}}$, $\mathcal{F}^{\mathcal{W}^M_{Level-2}}$, and $\mathcal{F}^{\mathcal{W}^M_{Level-3}}$.

2.4 Pre-trained vision-language model

We adopt QUILT-Net [13] as the VLM $\mathcal{V}$ for histopathology image analysis due to its superior performance compared to state-of-the-art VLMs such as CLIP and PLIP. In QUILT-Net, the visual encoder is constructed based on ViT-B/32 [4], which splits input images into 32x32 tiles and forwards through 12 self-attention layers with 12 heads. The text encoder inherits an architecture of GPT-2 [18] with 12 self-attention layers with 8 heads. These encoders are initialized with the weights of CLIP and then fine-tuned with QUILT-1M.

3 Experiments

3.1 Datasets

We employ four different public datasets to analyze the effectiveness of TQx. The first dataset, called **Colon** [19], is a colorectal cancer grading dataset, which

Table 1. Silhouette coefficients of clustering with visual embedding and text-based embedding from four different WoI pools.

Embedding	Colon [19]	WSSS4LUAD [20]	BACH [21]	Bladder [22]
Visual	0.13	0.13	0.10	0.17
Text - $\mathcal{W}^M_{Level-0}$	0.25	0.26	0.21	0.27
Text - $\mathcal{W}^M_{Level-1}$	0.27	0.30	0.25	0.25
Text - $\mathcal{W}^M_{Level-2}$	0.28	0.29	0.25	0.23
Text - $\mathcal{W}^M_{Level-3}$	0.28	0.30	0.24	0.20

includes 9,857 patches of size 512 x 512 with four classes: *benign* (BN), *well-differentiated cancer* (WD), *moderately differentiated cancer* (MD), and *poorly differentiated cancer* (PD). **WSSS4LUAD** [20] is the second dataset for lung cancer detection. We extracted 3,526 patches of size 224 $\times$ 224 that are labeled as *normal* (NOR) or *tumor* (TUM). The third dataset is **BACH** [21] that contains 58,539 images of size 1024 x 1024 for breast cancer staging with four categories: *normal* (NOR), *benign* (BN), *in situ carcinoma* (SITU), and *invasive carcinoma* (IVS). The last dataset **Bladder** [22] is designed for bladder cancer grading, comprising 14,258 patches of size 512 $\times$ 512 that are annotated as *normal* (NOR), *low-grade cancer* (LOW), and *high-grade cancer* (HIGH).

3.2 Analysis of text-based image representation

To investigate the effectiveness of the text-based image representation, we conduct two tasks: 1) Clustering and 2) Classification. Both tasks are evaluated for four aforementioned WoI pools. To compare between visual and text-based embeddings, we adopt a pair of visual and textual encoders from QUILT-Net [13]. For **clustering**, we employ Lloyd's K-Means method [23], where K is set to a number of classification classes. K-Means++ is used for cluster initialization, with the maximum number of iterations is set to 300. We evaluate the clustering quality using silhouette coefficients and further assess the results visually through t-SNE [24].

For **classification**, we construct a simple multi-layered perceptron with two fully-connected layers, a ReLU activation function, and a batch normalization layer. The classifier is trained for 300 epochs using Adam optimizer with a learning rate 0.01. For Colon, Bladder, and BACH, four evaluation metrics are employed: accuracy (Acc), accuracy cancer (Acc_c), macro-averaged F1 score ($F1$), and quadratic-weighted kappa score (K_w). For WSSS4LUAD, Acc, $F1$, precision (Pre), and recall (Rec) are used. The experiments are replicated with 50 different initialization seeds to calculate the mean and standard deviation of each evaluation metric.

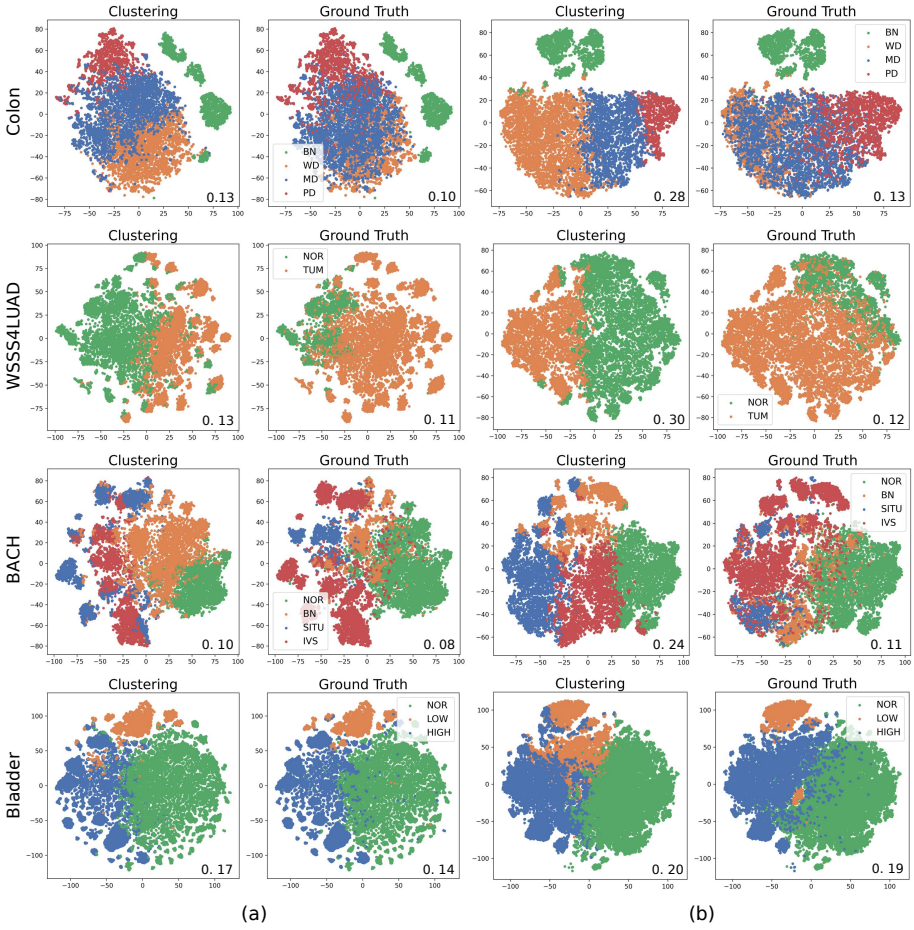

Fig. 2. Clustering results with $\mathcal{W}^{M}_{Level-3}$ (*Neoplastic Process*) of (a) visual embeddings and (b) text-based image embeddings. In the *Ground Truth* plots, samples are re-assigned to the clusters using the ground truth class labels. The bottom numbers show silhouette coefficients measuring how similar an embedding is to its own cluster.

4 Results and Discussion

4.1 Clustering

Fig. 2 visualizes the clustering results for $\mathcal{W}^{M}_{Level-3}$ using t-SNE, and Table 1 shows the silhouette coefficients of the four datasets. The results show that the text-based image embeddings well formed the clusters corresponding to the ground truth class labels across the datasets. It was striking that the silhouette coefficients of the text-based image embeddings were consistently larger than those of the visual embeddings such that $\geq$0.12, $\geq$0.13, $\geq$0.11, and $\geq$0.03 for

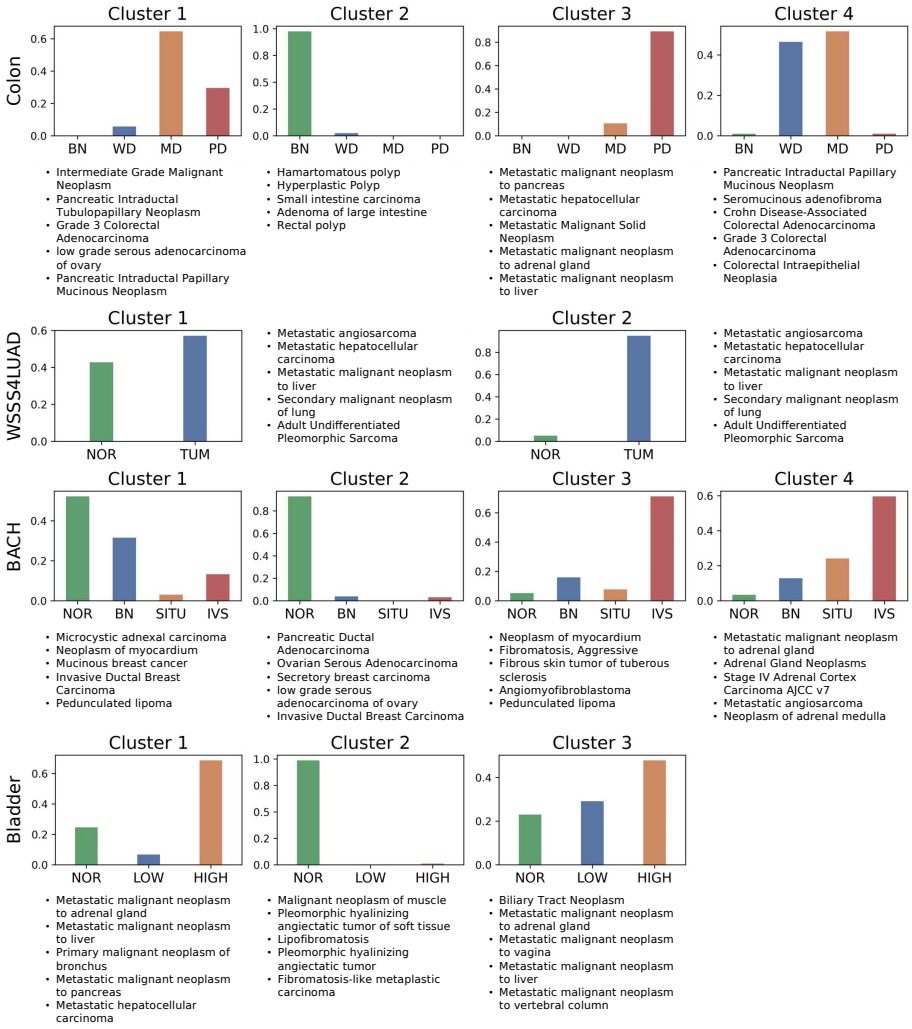

Fig. 3. The bar plots show the percentage of samples per class in each cluster, based on the clustering with $\mathcal{W}^M_{Level-3}$ (*Neoplastic Process*). Five keywords with the highest average ranks are shown next to the corresponding bar plot.

Colon, WSSS4LUAD, BACH, and Bladder, respectively. Among the text-based image embeddings, $\mathcal{F}^{\mathcal{W}^M_{Level-0}}$ generally obtained the worst results, whereas others were comparable to each other.

Fig. 3 depicts the percentage of data samples per class and the top-5 keywords within each cluster. We made the following observations. Each cluster was dominated by one or two categories that are related to each other; for example, Colon Cluster 4 was dominated by WD and MD, and TUM was predominant

in WSSS4LUAD Cluster 2. Non-cancerous samples were clearly separated from cancerous samples across the four datasets; for example, most of BN belonged to Cluster 2 for Colon, and most of NOR and BN resided in Cluster 1 and Cluster 2 for BACH, respectively. Furthermore, we analyzed the top-5 matching keywords per cluster. The keywords generally matched the type of dataset. For instance, in Colon, Cluster 2, dominated by BN, was associated with *Hamartomatous polyp* and *Hyperplastic polyp* that are non-cancerous, and other clusters (Cluster 1, 3, and 4), mainly containing cancers, were matched with cancer-related keywords such as *Grade 3 Colorectal Adenocarcinoma* and *Colorectal Intraepithelial Neoplasia*. For WSSS4LUAD, BACH, and Bladder, most of the clusters were shown to be highly relevant to cancer-related terms. Some of them are specific to the original organ, such as *Secondary malignant neoplasm of lung* for WSSS4LUAD and *Invasive Ductal Breast Carcinoma* for BACH. Some others were related to different organs; for example, *Pancreatic Ductal Adenocarcinoma* for BACH and *Metastatic malignant neoplasm to adrenal gland* for Bladder. Though cancer-related terms were prevalent, the clusters representing BN and/or NOR were often aligned with benign, slow-growing tumors such as *Pleomorphic hyalinizing angiectatic tumor* (slow-growing tumor) and *Lipofibromatosis* (benign soft tissue tumor).

4.2 Classification

The classification results on the four datasets are shown in Table 2. For Colon and BACH, the visual embeddings performed best, outperforming the text-based image embeddings over the four evaluation metrics in Colon and three evaluation metrics (*Acc*, Acc_c, and $F1$) in BACH. For WSSS4LUAD, $\mathcal{F}^{\mathcal{W}^M_{Level-3}}$ achieved the best performance in *Acc*, Acc_c, and $F1$. As for Bladder, the visual embeddings and text-based image embeddings were comparable to each other; the visual embeddings obtained the best *Acc* and K_w, while the highest Acc_c and $F1$ were attained by the text-based image embeddings ($\mathcal{F}^{\mathcal{W}^M_{Level-0}}$).

Among the four types of the text-based image embeddings, $\mathcal{F}^{\mathcal{W}^M_{Level-3}}$, in general, achieved the best performance and $\mathcal{F}^{\mathcal{W}^M_{Level-0}}$ was inferior to others except for Bladder. Hence, the more specific the WoI pool is, the better performance we tend to obtain. This indicates that the performance of the embeddings depends on the selection of the WoI pool, and the well-defined WoI pool could further improve the histopathology image analysis by instructing the specific patterns in histopathology images.

5 Conclusions

Herein, we propose TQx, a text-based quantitative and explainable histopathology image analysis framework that exploits a pre-trained VLM and a simple image-to-text retrieval. The text-based image embeddings, driven by TQx, can

Table 2. Classification results using visual and text-based image embeddings.

Embedding	Colon [19]				WSSS4LUAD [20]			
	Acc (%)	Acc_c (%)	*F*1	K_w	*Acc* (%)	*Pre*	*F*1	*Rec*
Vision	**81.7±1.8**	**74.6±2.5**	**0.778±0.014**	**0.913±0.008**	83.9±1.7	0.615±0.016	0.640±0.022	0.889±0.007
Text - $\mathcal{W}^M_{Level-0}$	77.0±2.7	68.7±4.2	0.729±0.019	0.895±0.011	87.3±0.4	0.639±0.003	0.681±0.005	**0.919±0.003**
Text - $\mathcal{W}^M_{Level-1}$	78.3±2.2	70.0±3.1	0.743±0.017	0.901±0.009	87.4±0.8	0.639±0.006	0.681±0.009	0.911±0.007
Text - $\mathcal{W}^M_{Level-2}$	78.5±1.7	70.2±2.3	0.748±0.012	0.902±0.007	87.0±1.2	0.637±0.009	0.677±0.015	0.913±0.009
Text - $\mathcal{W}^M_{Level-3}$	79.6±2.5	71.8±3.7	0.753±0.021	0.904±0.013	**89.3±0.5**	**0.645±0.005**	**0.691±0.007**	0.866±0.004
Embedding	BACH [21]				Bladder [22]			
	Acc (%)	Acc_c (%)	*F*1	K_w	*Acc* (%)	Acc_c (%)	*F*1	K_w
Vision	**75.1±1.7**	**64.8±2.4**	**0.660±0.020**	0.762±0.044	**86.1±0.9**	85.7±1.0	0.726±0.014	**0.749±0.016**
Text - $\mathcal{W}^M_{Level-0}$	72.7±0.6	58.1±0.8	0.603±0.016	0.765±0.008	85.2±0.8	**88.9±0.3**	**0.758±0.009**	0.716±0.013
Text - $\mathcal{W}^M_{Level-1}$	71.9±1.1	56.0±1.9	0.569±0.013	0.765±0.017	82.0±0.8	83.8±0.3	0.609±0.013	0.679±0.016
Text - $\mathcal{W}^M_{Level-2}$	72.3±1.4	57.0±2.1	0.574±0.011	0773±0.013	81.2±0.8	84.0±0.3	0.604±0.015	0.662±0.013
Text - $\mathcal{W}^M_{Level-3}$	73.1±2.1	58.6±1.3	0.564±0.017	**0.801±0.013**	78.2±0.4	88.3±0.4	0.683±0.010	0.594±0.005

be used for quantitative analysis and are directly associated with the histopathology terms, which are human-readable and -understandable without any post-processing or interpretation. The future study will further investigate the construction and optimization of the WoI pool and the application of the text-based image embeddings for other downstream tasks.

Acknowledgments. This study was supported by the National Research Foundation of Korea (NRF) (No. 2021R1A2C2014557) and by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (P0022543).

Disclosure of Interests. There are no competing interests.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
2. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. International Conference on Machine Learning (2021). https://doi.org/10.48550/arXiv.2104.00298
3. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Conference on Computer Vision and Pattern Recognition (2022). https://doi.org/10.48550/arXiv.2201.03545
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021). https://doi.org/10.48550/arXiv.2201.03545
5. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11999–12009 (2022). https://doi.org/10.1109/CVPR52688.2022.01170

6. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A.C., Li, Y.: Maxvit: Multi-axis vision transformer. In: European Conference on Computer Vision (2022). https://doi.org/10.48550/arXiv.2204.01697
7. Deng, A., Li, X., Hu, D., Wang, T., Xiong, H., Xu, C.Z.: Towards inadequately pre-trained models in transfer learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19397–19408 (October 2023)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
9. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
10. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I., Roth, D.: Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys **56**(2), 1–40 (2023)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
12. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**(9), 2307–2316 (Sep 2023). https://doi.org/10.1038/s41591-023-02504-3, https://doi.org/10.1038/s41591-023-02504-3
13. Ikezogwo, W.O., Seyfioglu, M.S., Ghezloo, F., Geva, D.S.C., Mohammed, F.S., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=OL2JQoO0kq
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4904–4916. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/jia21b.html
15. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M., Naumann, T., Poon, H.: Large-scale domain-specific pretraining for biomedical vision-language processing (2023). https://doi.org/10.48550/ARXIV.2303.00915, https://arxiv.org/abs/2303.00915
16. Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., Fung, P.: Enabling multimodal generation on clip via vision-language knowledge distillation. arXiv preprint arXiv:2203.06386 (2022)
17. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Res. **32**, 267–270 (2004)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
19. Vuong, T.T.L., Kim, K., Song, B., Kwak, J.T.: Joint categorical and ordinal learning for cancer grading in pathology images. Medical Image Analysis **73**, 102206 (2021). https://doi.org/10.1016/j.media.2021.102206, https://www.sciencedirect.com/science/article/pii/S1361841521002516

20. Han, C., Pan, X., Yan, L., Lin, H., Li, B., Yao, S., Lv, S., Shi, Z., Mai, J., Lin, J., Zhao, B., Xu, Z., Wang, Z., Wang, Y., Zhang, Y., Wang, H., Zhu, C., Lin, C., Mao, L., Wu, M., Duan, L., Zhu, J., Hu, D., Fang, Z., Chen, Y., Zhang, Y., Li, Y., Zou, Y., Yu, Y., Li, X., Li, H., Cui, Y., Han, G., Xu, Y., Xu, J., Yang, H., Li, C., Liu, Z., Lu, C., Chen, X., Liang, C., Zhang, Q., Liu, Z.: Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. arXiv (2022)
21. Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., Aguiar, P.: Bach: Grand challenge on breast cancer histology images. Medical Image Analysis **56**, 122–139 (2019). https://doi.org/https://doi.org/10.1016/j.media.2019.05.010, https://www.sciencedirect.com/science/article/pii/S1361841518307941
22. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., Ahmad, N., Khalil, F.K., Dickinson, S.I., Shi, X., Liu, F., Su, H., Cai, J., Yang, L.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**(5), 236–245 (May 2019). https://doi.org/10.1038/s42256-019-0052-1, https://doi.org/10.1038/s42256-019-0052-1
23. Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489
24. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html