Histopathology Cross-Modal Retrieval based on Dual-Transformer Network

1st Dingyi Hu, 4th Fengying Xie, 5th Zhiguo Jiang
Image Processing Center,
Beijing Advanced Innovation
Center for Biomedical Engineering,
Beihang University, Beijing, China

2nd Yushan Zheng
School of Engineering Medicine,
Beijing Advanced Innovation
Center for Biomedical Engineering,
Beihang University, Beijing, China
yszheng@buaa.edu.cn

3rd Jun Shi
School of Software,
Hefei University of Technology,
Hefei, China.

***Abstract*—Computer-aided cancer diagnosis (CAD) methods based on the histopathological images have achieved great development. The content-based whole slide image (WSI) retrieval is one of the important application that can search for the informative data to assist clinical diagnosis. It is notable that the current retrieval system are mainly developed based on the image content and image labels. The diagnosis report for the WSIs given by the pathologists are also valuable data, but have not yet been adequately considered in modeling. In this paper, we propose a cross-modal retrieval framework based on histopathology WSIs and diagnosis report, which can simultaneously achieve four retrieval tasks for histopathology database across WSIs and diagnosis reports. The compact binary features from both WSIs and diagnosis reports are first extracted, and then built in a common vision-language semantic feature space by the constraint of the designed cross hashing loss function. The method was verified on a gastric histopathology dataset that contains 932 gastric cases with 4 lesion categories. Experimental results have demonstrated the effectiveness of the proposed method in the cross-modal retrieval tasks for digital pathology system.**

***Index Terms*—Histopathology image analysis, Cross-modal retrieval, CAD**

I. Introduction

In recent years, Computer-aided cancer diagnosis methods based on histopathology whole slide image (WSI) analysis [1]–[5] have been widely developed and proven promising in clinic diagnosis. Content-based image retrieval [5], [6] (CBIR) for histopathology WSI is an emerging technique in this domain. It is designed to analyze the query data and search the database for effective diagnosis information. The CBIR methods based on convolutional neural network (CNN) [**?**], [7] have become popular for the advantages in dealing with large-scale data. However, the size variety of the query image presented challenge to the CNN-based methods that are designed for images in fixed size. To solve the problem, graph neural networks (GNN) was applied in histopathology image analyze and CBIR [8], [9], which can extract structural features for the tissue and can break the limitation of the query image size. However, the current CBIR models for histopathology images are built based on image-level labels, which makes the present methods suffer from the problem of feature dimension collapse, information redundancy [10], etc. It is notable that the clinical diagnosis reports made by the pathologists for the histopathology WSIs are informative to the WSIs. This information is potential to improve the performance of the CBIR but has not yet been properly utilized in the previous methods.

With the development of computer vision and natural language processing techniques, the cross-modal retrieval methods [11], [12] are well studied. Cao et al. [13] proposed a hybrid networks based on deep visual-semantic quantification for efficient image retrieval. Liu et al. [14] designed a Transformer-based model that leverages rich pre-trained vision-and-language (V&L) knowledge to build a semantic vision-language space, where the database items in different modality can be indexed for cross-modal retrieval. Several researches [15]–[18] proved the diagnosis report are useful for the medical image analysis. Xue et al. [18] proposed a novel generative model which incorporates the CNN with the recurrent module of long short-term memory (LSTM), to generates a complete radiology report. Wu et al. [17] employed a light-weighted CNN to extract histopathological image features and designed a attention LSTM structure to generate both target text and attention mask. The previous methods have shown the effectiveness of the diagnosis texts in the improvement of medical image analysis performance.

In this paper, we proposed a novel framework for cross-modal retrieval of histopathology whole slide images and diagnosis reports with a deep cross-modal hashing network. As shown in Fig.1, in the data collection and model training stage, we collect the WSIs with diagnosis reports and train the proposed network. Next the compact binary feature is extracted to form the multi-modal hash table. In the retrieval stage, the WSIs or reports query are input to the trained model to obtain the corresponding hash feature. Finally, the relevant WSIs and reports are returned from the database based on hash search.

The contribution of our work can be summarized into three points.

(1) We propose a novel cross-modal retrieval framework

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61901018, 62171007, and 61906058), partly supported by the Anhui Provincial Natural Science Foundation (Grant No. 1908085MF210), and partly supported by the Fundamental Research Funds for the Central Universities of China (Grant No. JZ2020YYPY0093).

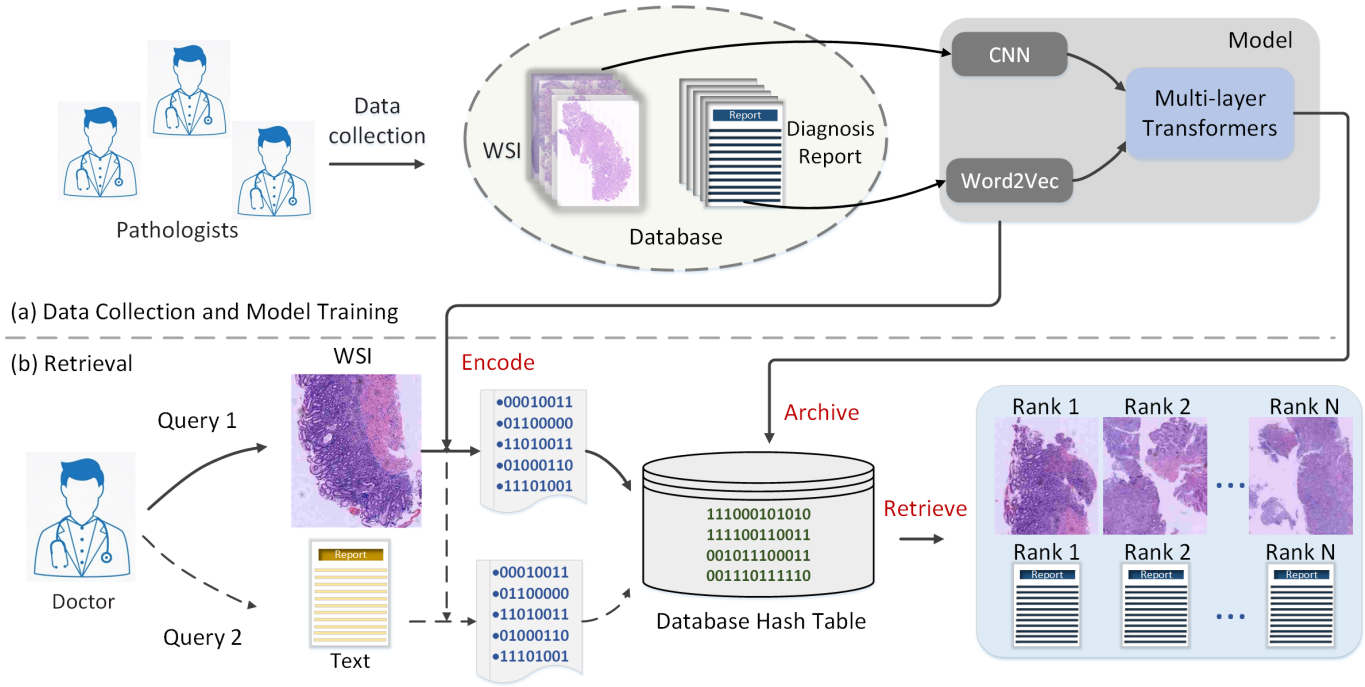


Fig. 1. The flowchart of proposed retrieval methods. (a) is the overview of the data collection and model training, where the diagnosis report and WSIs are meanwhile input to the cross-modal hash encoding network, and the binary codes for the two modality data are obtained as index for retrieval. (b) illustrates two queries from different modality, where then relevant cases involving in similar images and texts are returned from the database.

for digital pathology that allows the doctors to retrieval the database using either WSI and text input. To our knowledge, we are the first to tackle the cross-modal retrieval problem of histopathology WSIs and diagnosis reports. It significantly improves the convenience and applicability of the proposed retrieval framework compared to the previous methods.

(2) A dual branch network is proposed to learn the semantic information from both histopathology images and report texts under the constraint of a designed cross hashing loss function. The training of the proposed models depends on the diagnosis report and WSI category information, rather than the hand-craft annotations for the WSI, which determines it is data-efficient and appropriate for large-scale database.

(3) Comprehensive experiments are conducted to verify the proposed method and compared it with the related method [9], [19], [20] on a gastric dataset with 932 cases. The experiments demonstrate the effectiveness of the proposed cross-modal retrieval framework.

II. Method

In this section, we expatiate the proposed cross-modal retrieval method in three phases, Cross-modal encoding network, Cross hashing loss function, and train and application.

A. Cross-modal encoding network

The main task of the network is to simultaneously extract the global image representation of a WSI and the semantic representation from the diagnosis reports, and then to project the representations from the two branches into a common semantic feature space. To achieve this, we build a two-branch network, as shown in Fig.2.

In the first branch, we utilize the ResNet50 [21] structure pre-trained on ImageNet [22] to extract the local features of WSIs. To tackle the problem of the gigapixel property of WSIs, we divided the WSI into non-overlapping tiles in size of 224×224 and fed them into the ResNet50 to extract the features. Specifically, assuming $\{I_i|i = 1, ..., N\}, I_i \in \mathcal{R}^{224\times224\times3}$ are N image tiles divided from a WSI, we obtain the corresponding feature $\{p_i|i = 1, ..., N\}, p_i \in \mathcal{R}^D$, where D is the dimension of the last connection layer of the pre-trained CNN. Then we apply a fully connected layer and a normalization layer before the transformer module, to further reduce the feature dimension from D to d. Finally, we randomly select $\hat{N}$ samples from all the tiles, to make a balance of the computational complexity of the accuracy of the self-attention module.

In the second branch, we adopt Word2Vec [23] as our text model to transfer the words in diagnosis report into vectors. The branch model is trained to make the features with similar semantics have a relatively close distance in the feature space, which is the essential for information retrieval. Specifically, assuming $\{w_j|j = 1, ..., M\}$ with M representing the number of words in a diagnosis report, we applied the Word2Vec model to obtain corresponding vectors $\{v_j|j = 1, ..., M\}$. As shown in Fig.2, we symmetrically appended a fully connected layer and a normalization layer to the Word2Vec model. Finally, we can obtain the d dimension text features, which is represented as $\{t_j|j = 1, ..., M\}, t_j \in \mathcal{R}^d$.

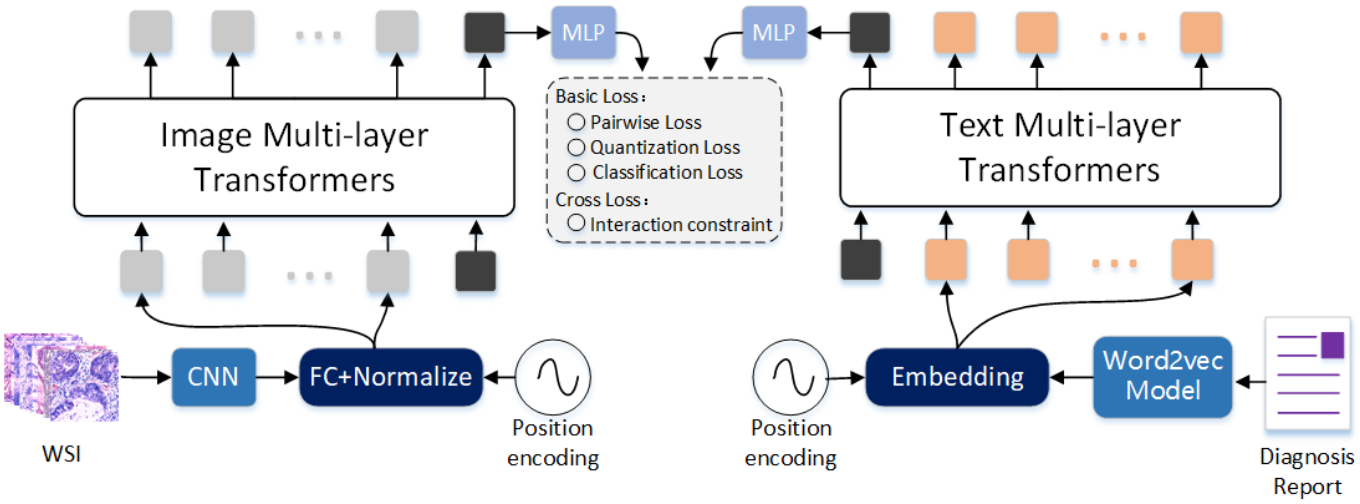


Fig. 2. The overview of proposed retrieval methods. The WSI and diagnosis report are transformed into feature tokens and input into a symmetrical Transformer-based model respectively.

Multi-layer Transformer is applied to build the two branches. The text and image parts shares similar network, as shown in Fig.2. The spatial position information of image tiles and the relative position of words have important effect to global features. Hence, we build the position embedding for the image features and text vectors referring to [20]. Then, we concatenate a trainable class token $S^0_{cls} \in R^d$ to the feature sequence for the information fusion of all the tokens, and outputs the case-level representations. Each Transformer block contains a multi-head attention (MHA), a feed-forward network (FFN) and 2 Layernorm (LN) and residual connection, which is the same as Transformer [20].

B. Cross-modal hashing loss

We build a linear layer with tanh activation function on the end of each branch to achieve the hash encoding, which renders the framework a fast retrieval speed and a low database storage. The hashing module can be defined as $b = sgn(z), z = MLP(S^L_{cls})$, where $sgn(\cdot)$ denotes the sign function, $sgn(x) = 1$ if $x > 0$, and $sgn(x) = -1$ otherwise. Then, we design a hybrid hashing loss function to guide the cross-modal information conjunction. The loss function involves in three parts.

The first loss function is the cross-modal loss that is used to project the representations from the image and text to the same feature space, which is the main means to achieve the cross-modal retrieval. Supposing z^i_n and z^t_n denotes the binary-like codes out of the image-branch and the text-branch for the n-th sample, the cross-modal loss is defined as

$$L_x = \frac{1}{S}\sum_{m=1}^{S}\frac{1}{S}\sum_{n=1}^{S}\left|z^i_m \cdot z^t_n - d_h y_{mn}\right|, \tag{1}$$

where d_h denotes the length of the hash code, y_{mn} indicates the relation of the m-th and n-th samples, where $y_{mn} = 1$ denotes m-th and n-th samples are relevant, and $y_{mn} = 0$ otherwise. For the proposed application of cross-modal retrieval for digital pathology, we define two samples (including WSIs and texts) are relevant if they are from the cases sharing the same lesion type, and irrelevant, otherwise. The second part is the pairwise loss that is used to control the distances of the WSI representations referring to their lesion types, which is defined as

$$L_p = \frac{1}{S}\sum_{m=1}^{S}\frac{1}{S}\sum_{n=1}^{S}\left|z^i_m \cdot z^i_n - d_h y_{mn}\right|, \tag{2}$$

The Third loss function is the quantization loss that pushes the outputs of the hashing layers toward binary codes.

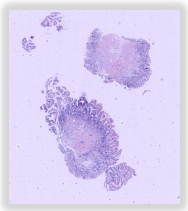
$$L_b = \frac{1}{S}\sum_{m=1}^{S}\left(\left|sgn(z^i_n) - z^i_n\right| + \left|sgn(z^t_n) - z^t_n\right|\right), \tag{3}$$

In addition, the image content plays the major role in the proposed WSI retrieval framework. In order to further enhance the performance of image-to-image retrieval, we make full use of the WSI category labels and append a classification loss function L_c in the training. Specifically, a FC layer with a softmax function is also built on the classification token of the image branch, besides the hashing layer, to produce the cross-entropy loss, which is represented as L_c. Finally, the losses used to train the network is combined as:

$$L = L_x + \alpha L_p + \beta L_b + \gamma L_c, \tag{4}$$

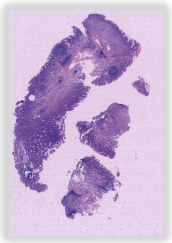
with α, β, and γ controls the weights.

(a) Low-grade intraepithelial neoplasia (LGIN)



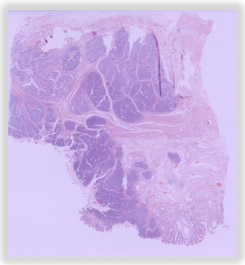
1. (Posterior wall of gastric antrum) mucosal chronic active inflammation with intestinal metaplasia and a small amount of bleeding in individual glands, as well as a small amount of inflammatory exudation, necrosis, and low-grade intraepithelial neoplasia of individual glands.
2. (Posterior wall of the anterior pyloric area of the stomach) chronic active inflammation of the mucosa with a small amount of bleeding, and a small amount of inflammatory exudation.
Helicobacter pylori (HP): (positive, +)

(b) High-grade intraepithelial neoplasia (HGIN)



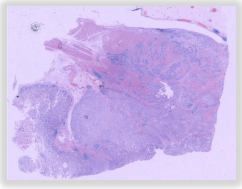
(Greater curvature of the pylorus) examined chronic inflammation of gastric mucosa with acute active inflammation, superficial erosion, inflammatory exudation, and focal glandular high-grade intraepithelial neoplasia.
(It is recommended to re-examine after treatment, and re-examine materials if necessary!)
Immunohistochemical labeling results (6579):
LCA(-), P53(-), Ki-67(50%+), Villin(Focus+), AE1/AE3(+).

(c) Adenocarcinoma (A)



(full stomach + anastomosis)
1. Tumor situation: (1) Histological type and grading: ulcerated moderately differentiated adenocarcinoma on the lesser curvature of gastric cardia.
2. Situation of incision margins: There was no cancer invasion in the annular incision margins at the upper and lower ends of the surgical specimen (0.8cm and 13.5cm away from the ulcer respectively after fixation) and the inspection (anastomosis).
3. Lymph node status: 15 lymph nodes were found on the greater curvature of the stomach, 25 lymph nodes were found on the lesser curvature, a total of 40 lymph nodes, and no cancer metastasis was found (0+/40).
Pathological diagnosis: ulcerated moderately differentiated adenocarcinoma of the lesser curvature of the gastric cardia (pathological TNM stage: pT2N0Mx).
Tumor tissue immunohistochemical markers (6388) results: Her-2 (-), P53 (90%+), Ki-67 (30%+), EGFR (-), VEGF (-), Syn (-), CgA (-), CD56 (-). "

(d) Signet-ring cell carcinoma (SRCC)



(total gastrectomy specimen)
1. Tumor situation: (1) Histological type and grading: gastric corpus lesser curvature ulcerated moderately-poorly differentiated adenocarcinoma.
(2) Ulcer size: 4.5 × 2.8 cm, the cancer tissue invaded the whole thickness of the stomach wall and reached the extraserous soft tissue.
(3) Cancer invasion was seen in nerves, and no cancer invasion was seen in vessels.
2. Incision margin situation: the upper and lower ends of the surgical specimen are annular incision margins (1cm and 7cm away from the ulcer respectively after fixation)

Fig. 3. Instances of cases in the gastric database, where the gastric lesions category are provided on the top left of each sample.

C. Train and application

The entire model was trained end-to-end by the Adam optimizer with primary learning rate 5×10^{-5} which was decrease in cosine function. The parameters of the Transformer block was randomly initialized. We empirically set the weights of the loss functions $\alpha = 0.1, \beta = 0.5, \gamma = 0.5$ referring to experimental results that the final performance is insensitive to the three hyper-parameters. After training, the two branches can be separately applied to encode the WSI or text. Then, the retrieval can be achieved by measuring the hamming distance between the query binary code and those in the database. Because the representations of the two modality data are projected to the same indexi

ng space through the cross-modal training, the doctors are allowed to retrieve the database using either WSI and text, or use both modalities. It significantly improves the convenience and applicability of the proposed retrieval framework compared to the previous methods.

TABLE I
PERFORMANCE OF 4 RETRIEVAL WSI TASKS IN THE DATABASE, WHERE THE BEST VALUES ARE SHOWN IN BLOD, THE "-" MEANS THE METHOD CAN NOT WORK ON THE CORRESPONDING TASK.

Methods	Image-to-Image		Image-to-Text		Text-to-Text		Text-to-Image	
	P@5	MAP	P@5	MAP	P@5	MAP	P@5	MAP
Transformer [20]	-	-	-	-	0.861	0.886	-	-
ViT [19]	0.763	0.786	-	-	-	-	-	-
DRA-Net [9]	0.810	0.836	-	-	-	-	-	-
proposed w/o cls	0.809	0.820	0.785	0.805	**0.885**	**0.903**	**0.910**	**0.923**
proposed	**0.824**	**0.838**	**0.802**	**0.834**	0.873	0.892	0.905	0.917

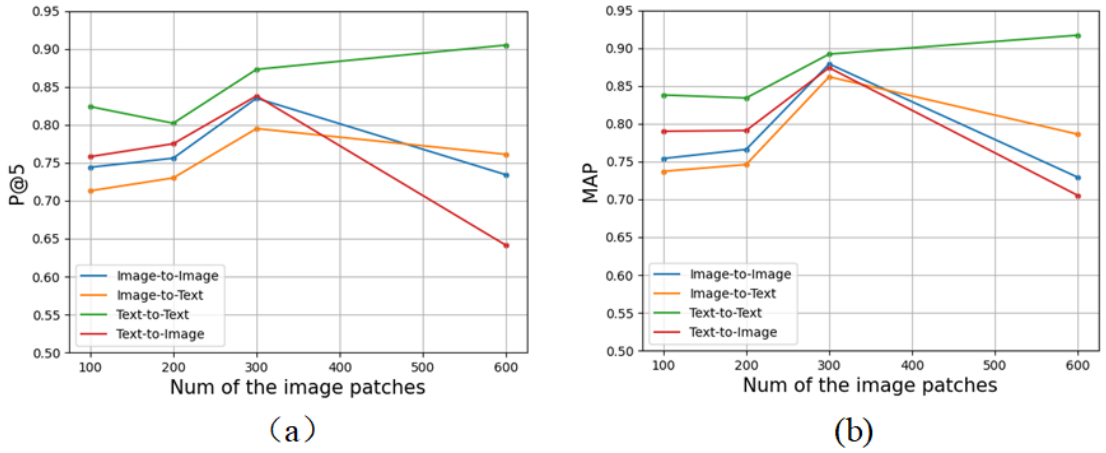


Fig. 4. The retrieval performance under different parameters. (a)The relationship of different image patch number and top 5 prediction. (b)The MAP performance under different image patch number

III. EXPERIMENT

A. Experimental settings

To study the effectiveness of the proposed method, we collect 932 H&E stained WSIs from gastric patients in 4 category of gastric lesions, including Low-grade intraepithelial neoplasia (LGIN), High-grade intraepithelial neoplasia (HGIN), Adenocarcinoma (A.), and Signet-ring cell carcinoma (SRCC). Each WSI was collected from one case that contains the corresponding diagnosis report by the well-trained pathologists. The WSIs were scanned under lenses of $20\times$ (the resolution is 0.48 μm/pixel). Several samples from the dataset are presented in Fig.3. In the experiment, 132 WSIs were randomly sampled from the dataset as the test set for query usage and the remainder were used to train and construct database.

Each word in the report was compress into a 50 dimensional vector by the trained skip-gram model [23]. The dimension of the image feature is 2048 for the structure of ResNet50. The image and text features were project into a 256-dimensional feature space and the length of the hashing code was set $d_h = 64$. We stacked 6 Transformer blocks in the image encoding branch and 3 blocks in text encoding branch. All the algorithms were implemented in python with PyTorch and run on a computer cluster with 4 available GPUs of Nvidia Geforce 3090.

B. Result and Discussion

We evaluated the proposed method in the following four retrieval tasks: Image-to-Image, Image-to-Text, Text-to-Text and Text-to-Image, and compared it with typical retrieval methods for text and images including Transformer [20], ViT [19], and DRA-Net [9]. Besides we also evaluated the model trained without the classification loss function. The precision of top-N returned instances (P@N) and mean average precision (MAP) were used to as the metrics. The results are shown in Table I.

Compared with the traditional single modal retrieval method [19], [20], the proposed method achieved higher accuracy. Especially on the Image-to-Image task, our method achieved a P@5 of 0.824 and an MAP of 0.838, which is 0.046 and 0.052 higher than ViT [19], respectively. The results have demonstrated that the proposed method successfully leverages the information from the text into the learning of image content. DRA-Net [9] applies the browsing path information of the diagnosis pathologist on the WSI as the auxiliary supervision besides the WSI label. The results in Rows

3&5 show that our method performs better than DRA-Net. It indicates that the diagnosis report contains more definite semantic description and is more informative than browsing path in building the retrieval system. Besides, the cross-modal supervision enables our method to be applied in another 3 retrieval tasks. Comparing the methods in Row 4&5, we can see that the loss with L_c achieves better performance on the tasks that take the images as input, and is contrast otherwise. Therefore, we recommend to use the loss with L_c when aiming at building image-input retrieval system, and to disable L_c in the loss when building multi-modal retrieval system for multiple retrieval requirements.

1) Hyper-parameter Sensitivity Analysis.: In this experiment, we verify the influence of image patch number per WSI on the retrieval performance. The retrieval performance as a function of the patch number was drawn in Fig.4. It shows the overall performance increases when the patch number enlarges from 100 to 300. It indicates the correct encoding of the WSI content depends on an adequate amount of local information. However, the overall performance decrease when the number beyond 300. It may because the significant scale difference of the tokens for the WSI and text encoding branch, where large WSIs contain over 5,000 image patches but its diagnosis report only includes 30 words. Excessive image patches on the image encoding branch would noisy the semantics mining on the text encoding branch.

IV. Conclusion

In this paper, we proposed a novel deep cross-modal retrieval hashing framework that realized four kinds of retrieval tasks across WSIs and diagnosis reports. The proposed method is the first attempt to tackle cross-modal retrieval problem of histopathological WSIs and diagnosis reports. The proposed cross-modal retrieval network is trained based on WSI and the information from the diagnostic report without the pathologists' hand-craft annotations, which determines it can be applied on large-scale digital pathology database. The proposed method was evaluate on a large-scale gastric dataset. The results have demonstrated that our method is effective for cross-modal retrieval tasks for digital pathology system.

References

[1] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[2] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[3] T. C. Hollon, B. Pandian, A. R. Adapa, E. Urias, A. V. Save, S. S. S. Khalsa, D. G. Eichberg, R. S. D'Amico, Z. U. Farooq, S. Lewis *et al.*, "Near real-time intraoperative brain tumor diagnosis using stimulated raman histology and deep neural networks," *Nature medicine*, vol. 26, no. 1, pp. 52–58, 2020.

[4] S. Kalra, H. R. Tizhoosh, S. Shah, C. Choi, S. Damaskinos, A. Safarpoor, S. Shafiei, M. Babaie, P. Diamandis, C. J. Campbell *et al.*, "Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–15, 2020.

[5] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, Y. Ma, H. Shi, and Y. Zhao, "Histopathological whole slide image analysis using context-based cbir," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1641–1652, 2018.

[6] M. Sapkota, X. Shi, F. Xing, and L. Yang, "Deep convolutional hashing for low-dimensional binary embedding of histopathological images," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 805–816, 2018.

[7] X. Shi, M. Sapkota, F. Xing, F. Liu, L. Cui, and L. Yang, "Pairwise based deep ranking hashing for histopathology image classification and retrieval," *Pattern Recognition*, vol. 81, pp. 14–22, 2018.

[8] M. Wojciechowska, S. Malacrino, N. Garcia Martin, H. Fehri, and J. Rittscher, "Early detection of liver fibrosis using graph convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 217–226.

[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, no. 3, pp. 1090–1103, 2020.

[10] D. Hu, Y. Zheng, H. Zhang, S. Sun, F. Xie, J. Shi, and Z. Jiang, "Informative retrieval framework for histopathology whole slides images based on deep hashing network," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 244–248.

[11] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] Y. Cao, M. Long, J. Wang, and S. Liu, "Deep visual-semantic quantization for efficient image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1328–1337.

[14] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2125–2134.

[15] Y. Chai, H. Liu, and J. Xu, "Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models," *Knowledge-Based Systems*, vol. 161, pp. 147–156, 2018.

[16] J. Tian, C. Li, Z. Shi, and F. Xu, "A diagnostic report generator from ct volumes on liver tumor with semi-supervised attention mechanism," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 702–710.

[17] F. Wu, H. Yang, L. Peng, Z. Lian, M. Li, G. Qu, S. Jiang, and Y. Han, "Agnet: Automatic generation network for skin imaging reports," *Computers in biology and medicine*, p. 105037, 2021.

[18] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 457–466.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.