WsiCaption: Multiple Instance Generation of Pathology Reports for Gigapixel Whole-Slide Images

Pingyi Chen[1,2,3], Honglin Li[1,2,3], Chenglu Zhu[2,3], Sunyi Zheng[2,3], Zhongyi Shui[1,2,3], and Lin Yang[2,3]

[1] Zhejiang University
[2] Research Center for Industries of the Future, Westlake University
[3] School of Engineering, Westlake University.
{chenpingyi,yanglin}@westlake.edu.cn

Abstract. Whole slide images are the foundation of digital pathology for the diagnosis and treatment of carcinomas. Writing pathology reports is laborious and error-prone for inexperienced pathologists. To reduce the workload and improve clinical automation, we investigate how to generate pathology reports given whole slide images. On the data end, we curated the largest WSI-text dataset (PathText). In specific, we collected nearly 10000 high-quality WSI-text pairs for visual-language models by recognizing and cleaning pathology reports which narrate diagnostic slides in TCGA. On the model end, we propose the multiple instance generative model (MI-Gen) which can produce pathology reports for gigapixel WSIs. We benchmark our model on the largest subset of PathText. Experimental results show our model can generate pathology reports which contain multiple clinical clues and achieve competitive performance on certain slide-level tasks. We observe that simple semantic extraction from the pathology reports can achieve the best performance (0.838 of F1 score) on BRCA subtyping surpassing previous state-of-the-art approaches. Our collected dataset and related code are available at https://github.com/cpystan/Wsi-Caption.

Keywords: Whole Slide Images · Image Caption · Visual-language Learning.

1 Introduction

Whole-slide images (WSI) based diagnostic pathology is the foundation and gold standard for the diagnosis of carcinoma. Due to the enormous size and large amount of heterogeneous information that exists in WSIs, the reading and interpretation of the slide usually necessitates specialized pathologists. Recently, with the advancement of computer-aided methods of WSIs, computational pathology has achieved remarkable success and, assisted with deep learning, some can even outperform experienced pathologists in certain tasks [24,29,4,5]. These advanced methods have largely improved the automation of the pathological reading workflow, especially for those less-experienced pathologists in rural areas [30].

In spite of the "clinical-grade" performance of these computational pathology approaches, pathologists still need to organize the findings and write textual reports for each slide. Hundreds to thousands of WSIs need to be summarized in text in the pathology departments every day [9]. The automation of diagnostic reports can largely reduce the workload of pathologists. Furthermore, the content of pathology reports usually includes abundant diagnostic results [3]. Therefore, it motivates us to take a step forward to achieve the automatic generation of pathology reports. On the data end, the great advancement of computational pathology in the past years owes very much to the emergence of publicly available pathology datasets. Some researchers resort to books, articles, and webs [10,19,12] to obtain large-scale image-text pairs. However, their collected images are small patches and the corresponding texts are also limited to patch-level descriptions. Therefore, collecting high-quality WSI-text pairs is worth exploring and can boost the development of visual-language models in computational pathology.

We notice that TCGA includes scanning copies of pathology reports in the format of PDF[4]. But they are too long with redundant information and present in a complex structure. Therefore, we propose a pipeline to extract and clean pathological texts from TCGA, which can convert complex PDF files to concise WSI-text pairs with the assistance of large language models (LLM).

There are still challenges to achieving slide-level generation on the model end. Recent years have witnessed the boom of visual-language models in image captioning [27,28,23]. And in the medical area, the generation of radiology reports has been explored by several works [14,7,18]. However, it is unaffordable to directly process the WSIs with more than 10 gigapixels unless WSIs are resized or disentangled sacrificing much fine-grained information. To deal with this problem, we introduce a **M**ultiple **I**nstance **Gen**eration (MI-Gen) framework to achieve WSI-based report generation. It contains a visual extractor which encodes the non-overlap patches with the sliding window and a sequence-to-sequence generator to produce pathology texts. In this work, our contributions can be concluded as follows:

1. We propose a pipeline to curate high-quality WSI-text pairs from TCGA. The dataset (PathText) contains about ten thousand pairs which will be publicly accessible. It can potentially promote the development of visual-language models in pathology.
2. We design a multiple instance generation framework (MI-Gen) and benchmark our approach with various backbones in the dataset.
3. Our model can achieve a promising performance on certain slide-level tasks like tumor subtyping, surpassing previous state-of-the-art MIL methods.

2   Multiple Instance Generation: Problem Formulation

In the multiple instance generation framework, the huge image can be seen as a bag which contains a group of instances which are non-overlap patches with a
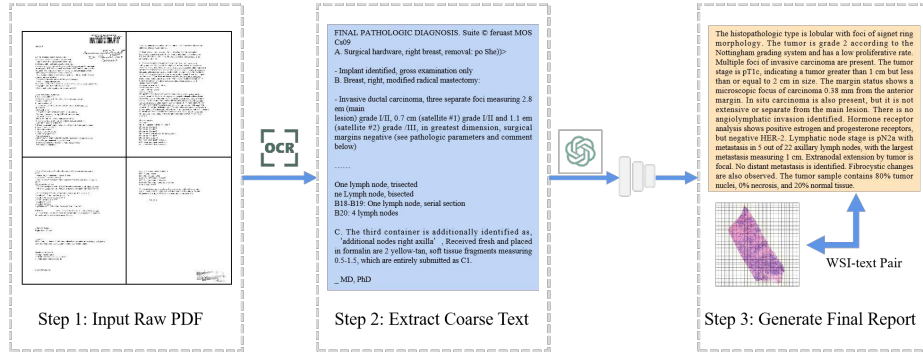
---

[4] https://portal.gdc.cancer.gov/

Fig. 1. The pipeline of extracting WSI-text pairs from TCGA.

much smaller resolution. Denote the cohort of instances as $Bag(X_i) = \{x_i^j\}_{j=1}^{M_i}$ where X_i is the i-th sample in the dataset and M_i is the sequence length which is determined by the patch size. Usually, when the patch size is 256, the sequence length M can be larger than ten thousand. A visual extractor h is adopted to extract image embeddings from the patches, denoted as $h(\{x_i^j\}) \in \mathbb{R}^{M_i \times l}$ where l is the embedding size.

A sequence-to-sequence model is incorporated to generate the target sequence $Y_i = \{y_i^j\}_{j=1}^{N_i}$ where N_i is the ground truth sequence length. Like previous text generative models, the encoder-decoder parameterized by ϕ is trained using language model objective which maximizes the sum of conditional possibilities of individual words in the sequence. Therefore, the negative log-likelihood (NLL) loss can be calculated as:

$$L = -\sum_i \sum_{t=1}^{N} log p_\theta(y_t | h(\{x_i^j\}), \{y_i^j\}_{j<t}; \phi). \tag{1}$$

The probability of j-th word is calculated based on the patch embeddings and the previous sequence $\{y_i^j\}_{j<t}$.

3   Method

3.1   PathText Construction

The first step in constructing PathText is to find out the diagnostic slides and their corresponding pathology reports in the TCGA, as shown in Fig. 1. The diagnostic slides from TCGA cover diverse disease types originating at different primary sites. The pathology reports in the format of PDF usually contain multiple pages. We use OCR methods to transform the files into text [25]. However, the text is still noisy because the report itself contains redundant information, and OCR sometimes generates garbled code. Therefore, we apply LLMs [2] to filter and summarize the report with the prompt of *"Please summarize the following*
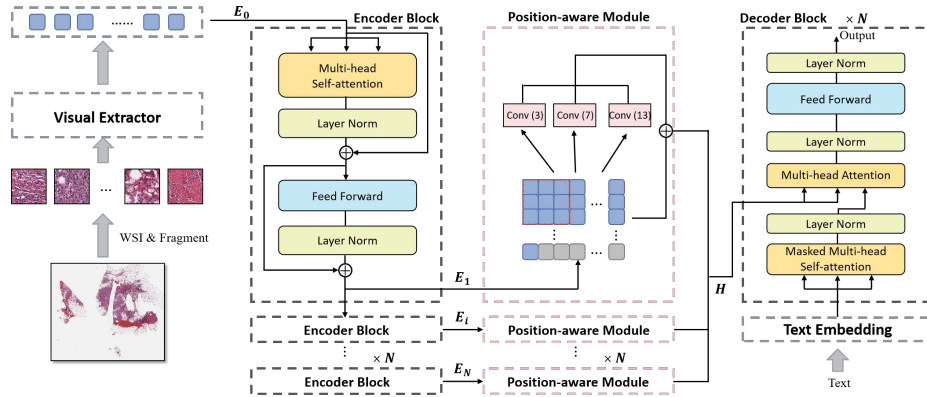
Fig. 2. The framework of our proposed model is comprised of a visual extractor and an encoder-decoder.

pathology report:". And to eliminate the influence of hallucination of LLMs, we manually annotate 88 pairs and train a classifier to remove those reports that are flawed.

PathText contains 9009 WSI-text pairs in total. Our collected text is distilled from clinical pathology reports owning well-aligned correspondence with WSIs and abundant pathological content. Our text is much longer than ARCH [10] (patch-level descriptions collected from books and articles). More details about PathText are demonstrated in the **Supplementary Materials**.

3.2   WSI-text Generation

As illustrated in Fig. 2, our proposed generative model consists of two parts: the visual extractor, and the encoder-decoder. We incorporate the hierarchical position-aware module into Transformer encoder layers so as to strengthen the model awareness of spatial information in WSIs. The details are described below.

Visual Extractor. Given a whole-slide image X_i with huge resolution, it is comprised of small patches $\{x_i^j\}_{j=1}^{M_i}$. It is unable to accumulate gradients in the visual extractor due to the large size of WSIs. Therefore, we apply pre-trained neural networks to extract features from these patches. The visual extractor is non-trainable during training. The initial image embeddings are denoted as E_0 which are fed to the subsequent modules.

Encoder-Dcoder. Transformer has shown remarkable success in the past few years due to its strong capability of modeling the long sequence and interaction among individual tokens with self-attention. Therefore, we adopt Transformer as our backbone. The Transformer encoder consists of N transformer layers. The embeddings are processed by the transformer layers iteratively as:

$$\boldsymbol{E}_i = f_i(\boldsymbol{E}_{i-1}) \tag{2}$$

where $f_i(\cdot)$ refers to the i-th transformer layer. Whereas, we propose hierarchical position-aware modules to aggregate the embeddings from different encoder layers. The position-aware modules are inserted after each encoder layer so that more abundant context information is captured. Therefore, the hidden state $\boldsymbol{H}$ for decoding can be formulated as:

$$\boldsymbol{H} = \sum_i PAM_i(\boldsymbol{E}_i). \tag{3}$$

Position-aware Module. It has been confirmed in [24] that convolutional layers improve position information awareness. Inspired by this, we incorporate a hierarchical position-aware module (PAM) into the encoding of image embeddings. Considering the varying sizes of tokens in WSIs, we also conduct padding so that the feature map can be reshaped for fitting convolutional layers. Convolution kernels of various sizes are adopted to capture heterogeneous spatial information.

In specific, denote the sequence of image embeddings as $\boldsymbol{E}_i = \{e_j\}_{j=1}^{M_i}$. Firstly, we pad the sequence until its length becomes a square number T_i. Then, the 1-D sequence can be reshaped into 2-D space $\boldsymbol{E}_i^{pad} \in \mathbb{R}^{\sqrt{T_i} \times \sqrt{T_i} \times l}$. We adopt several CNNs to process and aggregate the 2-D embeddings:

$$PAM_i(\boldsymbol{E}_i) = \sum Conv_i(\mathbf{E}_i^{pad}) + \mathbf{E}_i^{pad}. \tag{4}$$

Heterogeneous spatial information from the CNNs with varying kernel sizes is summed together. Finally, the hidden state returns to 1-D space as a sequence for decoding. Our PAMs are inserted hierarchically after each encoder block, encoding abundant spatial information at different depth, which improves the awareness of spatial descriptions in the generated reports.

4 Experiments and Results

4.1 Implementation Details

Datasets. We train and validate our generative model on the PathText (BRCA) which includes 845 pairs for training, 98 pairs for validating, and 98 pairs for testing. TCGA-BRCA contains 1041 whole slide images with the label of invasive ductal (IDC) or invasive lobular carcinoma (ILC). It also contains the results of Her2 testing. 10% of TCGA-BRCA is randomly selected for validation and inference.

Table 1. Quantitative results of pathology report generation on PathText (BRCA). Different combinations of visual extractors and encoder-decoders are present for comparison. BLEU-n indicates the BLEU score computed based on n-grams.

Visual Extractor & Pre-train	Encoder-Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	$Fact_{ent}$
ResNet&ImageNet	CNN-RNN[27]	0.334	0.209	0.122	0.074	0.137	0.248	0.396
	att-LSTM[28]	0.367	0.234	0.128	0.085	0.151	0.262	0.442
	vanilla Transformer [26]	0.395	0.230	0.135	0.089	0.145	0.254	0.453
	Mem-Transformer [7]	0.317	0.207	0.136	0.091	0.129	0.270	0.467
	Ours	**0.403**	**0.254**	**0.168**	**0.117**	**0.163**	**0.280**	**0.498**
ViT&ImageNet	CNN-RNN[27]	0.328	0.201	0.127	0.082	0.142	0.253	0.380
	att-LSTM[28]	0.341	0.211	0.132	0.083	0.145	0.265	0.425
	vanilla Transformer[26]	0.346	0.216	0.137	0.091	0.149	0.273	0.443
	Mem-Transformer[7]	0.332	0.216	0.144	0.100	0.147	0.268	0.449
	Ours	**0.380**	**0.252**	**0.169**	**0.110**	**0.157**	**0.279**	**0.474**
ViT&HIPT	CNN-RNN[27]	0.342	0.215	0.141	0.084	0.148	0.260	0.403
	att-LSTM[28]	0.372	0.230	0.135	0.090	0.150	0.269	0.466
	vanilla Transformer[26]	0.383	0.237	0.151	0.096	0.152	0.264	0.488
	Mem-Transformer[7]	0.344	0.218	0.150	0.103	0.151	0.268	0.501
	Ours	**0.446**	**0.286**	**0.183**	**0.120**	**0.171**	**0.271**	**0.532**

Model Setting. The number of encoder layers and decoder layers are both 3. For self-attention modules, the number of heads is 4 and the size of embeddings is 512. Three CNNs are adopted in the PAM with the kernel size of 3, 7, and 13 respectively. We use Adam with the learning rate of 1e-4 to optimize the model. The weight decay is 5e-5. We adopt beam search with the size of 3 as the sampling method. Our model is trained on four A100-80G GPUs.

4.2   Baselines

To benchmark our models, two kinds of visual models are chosen as our visual extractor: ResNet [11] which is composed of convolutional layers and ViT [8] which is based on Transformer. For the pre-training of visual extractors, we explore two strategies: 1) ImageNet (out-of-domain) pre-training on extensive natural images, and 2) hierarchical self-supervised learning with a pyramid transformer (HIPT) [6] on TCGA (in-domain).

  We also apply several other popular backbones. There are two backbones which decode by LSTMs: CNN-RNN [27] and att-LSTM [28]. We set the hidden size of these two LSTM-based models as 512 and the number of layers is 3. We also re-implemented two state-of-the-art Transformer-based models. The first is Vanilla Transformer which can be seen as the ablated version of our method with the same structure except for the position-aware module. The other is Mem-Transformer [7] which is specially designed for medical report generation by incorporating a memory mechanism in the decoder.

4.3   Results

As illustrated in Table 1, we adopt standard image captioning evaluation metrics to evaluate the generation performance: BLEU [22], METEOR [1], and ROUGE
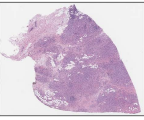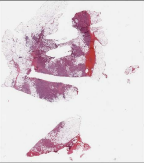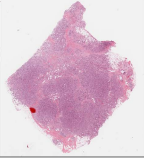
WSI	Ours	Vanilla Transformer	Ground-Truth
	The pathological slide shows that the patient underwent a lumpectomy for **invasive lobular carcinoma** in the breast. The tumor was identified as **2.6 cm** in size and located on the boundary of the outer quadrants. **There were no metastases in the axillary lymph nodes.** The tumor has a mitotic rate of 5 mitoses per 10 high-power fields. Hormone receptors estrogen and progesterone receptors were found in **over 75% of the tumor cells.** The **her2 protein was not detected** in the invasive cancer cells.	The pathological slide shows that there is no tumor present in the axillary lymph nodes. The axillary sentinel lymph node no tumor present. The axillary sentinel lymph node 3 no tumor present. The breast mastectomy shows **invasive lobular carcinoma.** ~~There is no tumor present.~~ There is no tumor present in the nipple or skin. Lobular carcinoma in situ lcis is present. The tumor size of the invasive carcinoma is less than 0.1 cm. ~~There is no tumor present.~~ The resection margins are free of tumor. There is no tumor present in the axillary lymph nodes. No metastases of tumor is present in the axillary lymph nodes.	The pathological slide describes a breast mastectomy with sentinel node biopsy. The breast specimen showed no remarkable findings. The breast specimen showed a **2.6 cm invasive lobular carcinoma** with no lymph node involvement. The tumor was graded as nottingham grade 1 and had no involvement of the skin or nipple. The tumor was **positive for estrogen and progesterone receptors** and **negative for her2.** The margin of excision was negative for carcinoma.
	The pathological slide indicates that there is no tumor present in the lymph nodes examined. However a biopsy of the breast shows the presence of **invasive ductal carcinoma** which is poorly differentiated. The invasive carcinoma measures **1.2 cm** in size and extends to within 5.0 mm of the nearest deep margin. There is also the presence of ductal carcinoma comprising a solid type with necrosis which **comprises less than 5% of the tumor area. All margins are free of tumor.** Additional immunohistochemical stains and molecular pathologic assays have been requested for further analysis.	The pathological slide indicates that the patient has **invasive ductal carcinoma** in the breast. The tumor has a nottingham grade 3 with a nuclear grade of 3 and **a ~~nuclear grade of there is no angiolymphatic invasion.~~** The surrounding breast tissue shows the presence of ductal carcinoma in situ. The tumor measures _1.0 cm_ in its largest dimension. There is no lymphovascular space invasion. **Resection margins are negative for tumor.** The invasive tumor has a ductal carcinoma in situ high nuclear grade and ductal carcinoma in situ is also present. The resection margins are negative for invasive tumor. The resection margins are negative for tumor except for the presence of atypical ductal hyperplasia and fibrocystic changes with associated calcifications.	The pathological slide shows that a local excision was performed for a mass measuring **1.3 x 1.3 x 1.2 cm.** The mass was diagnosed as **invasive ductal carcinoma** in situ ductal carcinoma in situ **comprising less than 5% of the tumor volume.** No angiolymphatic invasion is present . Biopsy site changes were observed. **All surgical resection margins are negative for tumor with a minimum tumor-free margin of 0.7 cm at the inferior margin.** In addition, the slides states that multiple axillary sentinel lymph nodes were examined and found to be negative for metastatic carcinoma.
	This is a pathological sliide for a patient who underwent a radical mastectomy for a tumor in the breast. The tumor was identified as **infiltrating ductal carcinoma** moderately differentiated. The size of the tumor was **2.1 x 1.8 x 1.8 cm.** No specific information was provided regarding the extent of the tumor or the margins. _The nottingham histologic score which assesses various characteristics of the tumor could not be determined based on the provided information._ There was no evidence of neo-adjuvant treatment or any additional pathologic findings .	The pathological slide indicates that the patient underwent a segmental mastectomy for a tumor in the breast. The tumor was identified as **infiltrating ductal carcinoma** moderately differentiated with **a ~~nottingham grade of there is no evidence of angiolymphatic invasion.~~** The surgical margins are negative for tumor . The tumor is estrogen and progesterone receptors are negative. The tumor is negative for her-2. The final diagnosis is infiltrating ductal carcinoma of the tumor and lymph node metastasis .	The patient underwent a lumpectomy for a tumor in the breast. The tumor measured **2.1 x 1.8 x 1.4 cm** and was identified as **infiltrating ductal carcinoma.** The tumor was moderately differentiated and the extent was not specified. There was no extracapsular invasion. The margins were uninvolved. The Nottingham Histologic Score indicated moderate tubule formation (score=2), marked nuclear pleomorphism (score=3), and a mitotic count of 10 to 20 mitoses per 10 high-power fields (score=2). The overall Nottingham Score was Grade II, indicating moderately differentiated cancer with a score of 6-7 points.

Fig. 3. Illustrations of pathology reports from our model, Vanilla Transformer, and ground-truth. The first column shows the thumbnails of the WSIs. The content that is consistent with the ground-truth is highlighted in bold. And the medical terms which are contradictory to ground-truth are underlined. Strikethrough text means there existing grammar errors.

[17]. We also adopt $Fact_{ent}$ [21] to measure the completeness and consistency of biological entities in the generated reports. We can observe that LSTM decoders perform relatively worse than transformer-based models. Mem-Transformer incorporates the relational memory to implicitly model similar patterns in different reports. However, it does not show a better performance than Vanilla Transformer. The potential reason might be that pathology reports in PathText are much longer than other medical image captioning datasets, thus having too heterogeneous structures to be memorized. In our model, we adopt hierarchical position-aware modules to capture spatial semantics in the WSI. This mechanism facilitates our model to achieve the best performance no matter what the visual extractor is.

Regarding visual extractors, ResNet and ViT do not show a large difference when pre-trained on ImageNet. But ViT pre-trained with HIPT demonstrates a significant improvement. It is not surprising because domain-aligned pre-training usually outperforms out-of-domain ImageNet pre-training in pathology [15].

Three samples of WSIs with their corresponding pathology report from different models are illustrated in Fig. 3. Our model generates more medical terms that are consistent with the ground-truth with less self-contradiction or grammar error. The pathology reports from ground-truth contain spatial information like tumor size ("2.6 cm", "1.3×1.3×1.2 cm"). Vanilla Transformer fails to give a tumor spatial description (the first and third WSI) or generate not precise

Table 2. The performance of slide-level tasks. Different MIL approaches are included for comparison. P, R, and F1 indicate precision, recall, and F1 score respectively.

Tasks / Method	Her2-prediction			Subtyping		
	P	R	F1	P	R	F1
Max-pooling	0.545	0.601	0.570	0.612	0.662	0.644
AB-MIL [13]	0.622	0.670	0.645	0.730	0.823	0.771
DS-MIL [16]	0.629	0.684	0.652	0.727	0.821	0.775
CLAM-SB [20]	0.605	0.668	0.632	0.781	0.879	0.823
TransMIL [24]	0.649	0.701	0.670	0.782	0.834	0.806
Semantic Extraction	**0.655**	**0.710**	**0.678**	**0.805**	0.865	**0.838**

results (the second WSI), which reflects its disability to capture spatial features. Our model provides the perfectly right descriptions in the first and second row ("2.6 cm", "less than 5% of the tumor area") and partially correct results for the others ("1.2 cm", "2.1×1.8×1.8 cm"). The potential reason why 3-D size can not be accurately predicted may be that the WSI can only provide 2-D spatial information. For the first WSI, though ground-truth only presents positive/negative results for hormone receptors, our model still has the tendency to generate spatial descriptions ("75% of the tumor cells", "not detected"). Although it is impossible to check their correctness, the corresponding binary classification of these spatial descriptions is exactly right. This comparison demonstrates that position-aware modules largely improve the spatial awareness of our model. More ablation studies can be found in **Supplementary Materials**.

Slide-level Tasks. To further verify the clinical performance of our model, we have compared WSI-related results (subtyping and Her2 prediction) which are included in the reports. We have observed that some generated reports contain the descriptions for carcinoma subtyping or Her2 prediction so that the corresponding result can be directly extracted from the text. Compared with multiple instance learning (MIL) approaches which are specially designed for WSI classification, we observe a more promising performance of our generative method, revealing the potential of our proposed visual-language learning method for computational pathology.

5   Conclusion

In this paper, we have shown the feasibility of pathology report generation. On the data end, we collected nearly 10000 WSI-text pairs by transforming PDF files in TCGA into concise and comprehensive pathology reports with the aid of LLMs. Our proposed PathText is the largest WSI-text dataset so far to the best of our knowledge and is going to be available to the public, which can promote visual-language learning in the pathology field. On the model end, we introduce MI-Gen as a generative model for WSI-level descriptions. By benchmarking different baselines on the subset of PathText, we also reveal the superiority of our

model in being aware of the spatial information among the WSIs and the promising performance on several slide-level tasks. In addition, other fields that utilize high-resolution images like remote sensing can be inspired by our work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Buckley, J.M., Coopey, S.B., Sharko, J., Polubriaginof, F., Drohan, B., Belli, A.K., Kim, E.M., Garber, J.E., Smith, B.L., Gadd, M.A., et al.: The feasibility of using natural language processing to extract clinical information from breast pathology reports. Journal of pathology informatics **3**(1), 23 (2012)
4. Chan, L., Hosseini, M.S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10662–10671 (2019)
5. Chen, P., Zhu, C., Shui, Z., Cai, J., Zheng, S., Zhang, S., Yang, L.: Exploring unsupervised cell recognition with prior self-activation maps. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 559–568. Springer Nature Switzerland, Cham (2023)
6. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022)
7. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Nov 2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Farahani, N., Parwani, A.V., Pantanowitz, L.: Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathology and Laboratory Medicine International pp. 23–33 (2015)
10. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16549–16559 (2021)

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
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 (2023)
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
14. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1240, https://aclanthology.org/P18-1240
15. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
16. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
17. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
18. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269. PMLR (2019)
19. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19764–19775 (2023)
20. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
21. Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. arXiv preprint arXiv:2010.10042 (2020)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
23. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017)
24. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
25. Smith, R.: An overview of the tesseract ocr engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
28. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
29. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18802–18812 (2022)
30. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**(5), 236–245 (2019)