Low-Shot Prompt Tuning for Multiple Instance Learning based Histology Classification

Philip Chikontwe[1], Myeongkyun Kang[2], Miguel Luna[2], Siwoo Nam[2], and Sang Hyun Park[2,3⋆]

[1] Department of Biomedical Informatics, Harvard Medical School, MA, USA
philip_chikontwe@hms.harvard.edu
[2] Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology, Daegu, Republic of Korea
shpark13135@dgist.ac.kr
[3] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, PA, USA

Abstract. In recent years, prompting pre-trained visual-language (VL) models has shown excellent generalization to various downstream tasks in both natural and medical images. However, VL models are sensitive to the choice of input text prompts, requiring careful selection of templates. Moreover, prompt tuning in the weakly supervised/multiple-instance (MIL) setting is fairly under-explored, especially in the field of computational pathology. In this work, we present a novel prompt tuning framework leveraging frozen VL encoders with (i) residual visual feature adaptation, and (ii) text-based context prompt optimization for whole slide image (WSI) level tasks *i.e.,* classification. In contrast with existing approaches using variants of attention-based instance pooling for slide-level representations, we propose synergistic prompt-based pooling of multiple instances as the weighted sum of learnable-context and slide features. By leveraging the mean learned-prompt vectors and pooled slide features, our design facilitates different slide-level tasks. Extensive experiments on public WSI benchmark datasets reveal significant gains over existing prompting methods, including standard baseline multiple instance learners.

Keywords: Histopathology · Multiple Instance Learning · Multi-Modal Learning · Weakly Supervised Learning

1 Introduction

Automated whole slide images (WSI) analysis using machine learning has been shown to mitigate tedious and laborious quantification, with the potential to serve as a secondary reader in clinical workflows [7, 9]. Despite this, algorithmic solutions are hindered by the nature of WSIs *i.e.*, extremely high resolution, stain variations across disease types and limited labeling. To address this, weakly supervised modeling of WSI with multiple instance learning (MIL) [1,5,16,19,23,25]

⋆ Corresponding author.

has become a standard approach for different tasks owing to patch-based learning to reduce compute heavy training, with instance pooling functions [11] adopted to obtain the final slide-level representation. While successful, learning with noisy labeled instances poses a challenge, especially in data-deficient settings. To solve this, self-supervised learning within and across WSI patch features has been explored to enable better transferability across tissue types [3, 6, 24] but may not scale to varied tissue types with limited samples.

Recently, the development of vision-language (VL) models [10,12,18,21] (*e.g.*, CLIP) using image-text pairs can not only learn aligned representations, but also enable improved zero-shot generalization on downstream tasks. As the models are supervised by natural language, images can be classified in open-vocabulary settings by placing the class name (*e.g.*, "A photo of a [CLASS]") in textual form. Note that fine-tuning VL models on downstream tasks is difficult and resource-intensive, and may damage learned features [14,26]. Thus, prompt tuning [17,22] is introduced to provide domain-specific context for downstream tasks *e.g.*, Context Optimization (CoOp) [30], Conditional Context Optimization (Co-CoOp) [29] and CoOp-GCE [27] enabled prompt tuning in CLIP by replacing context-words with learnable vectors and image-conditioning, with the latter an extension of CoOp for noisy learning. In more recent studies, unified prompt-tuning in both visual and textual encoders [4, 15] has been explored *e.g.*, CLIP-Adapter [8] introduces an alternative to prompting by fine-tuning with feature adapters in either visual or language branches reporting notable gains over CoOp variants.

In this work, we hypothesize that VL models can provide better separability for WSI classification despite learning with limited/noisy samples. First, as opposed to leveraging labeled samples with fine-tuning, prompt-tuning for WSI requires additional modules to pool instances for efficient learning. Thus, we propose to pool instances using learned context-vectors for improved instance-slide-text collaborative learning *i.e.*, we leverage K class-specific context vectors initialized with the text-model's special token (End Of Sequence: EOS) to weight the contribution of each instance in a WSI bag to obtain the slide-level representation. Second, inspired by the success of feature adaptation, we combine instance adaptation with learnable-prompting to simultaneously exploit knowledge learned in the VL model and new knowledge in the few-shot training examples. Final slide classification follows the standard similarity-based optimization, except we employ the average of K learned prompt vectors for scoring. In the context of this work, prompt tuning in MIL is fairly under-explored with limited works [18, 20, 28]. As opposed to visual-only prompting in PromptMIL [28] and using multiple prompt learning in *Qu et al.* [20], we explore a different design for text-based prompting on few-shot samples with visual adaptation.

The main contributions of this paper are as follows: (i) We study a challenging yet meaningful setting for prompt tuning *i.e.*, few-shot weakly supervised classification in histology, and (ii) introduce a novel prompt tuning framework using prompt optimization with context-driven instance pooling for slide-level classification and highlight the benefit of instance feature adaptation for im-

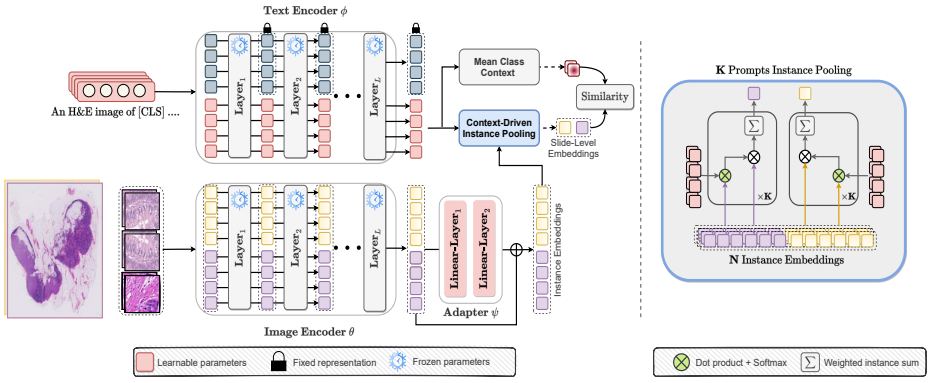

Fig. 1: **Proposed method**. For a given WSI, slide instances are fed to the visual encoder, followed by visual adaptation via $\mathcal{A}^v_\psi$. Note that different colored instances (purple/yellow) stem from different slides. On the other hand, the text-encoder takes text prompt input concatenated with new K learnable prompts per class. Before computing similarity scores, we perform instance-level pooling using the learnable projected K prompt vectors to obtain a single slide vector, and then use the mean class prompts for final slide-level prediction.

proved few-shot transfer. (iii) Extensive experiments and ablations on public datasets demonstrate the generalizability of our approach, achieving improved results in varying N-shot settings.

2 Method

Overview and Preliminaries. In this work, we consider a dataset of WSIs $X = \{X_i\}_{i=1}^N$ partitioned into non-overlapping patches $X_i = \{x_{i,j}, j = 1, 2, \ldots, n_i\}$ with varying n_i per WSI. Each X_i is associated with a slide-level label $Y_i = \{0, 1\}$, where $i = \{1, 2, \ldots, N\}$, respectively. Based on the standard MIL assumption [1], instances are associated with slide labels as follows: $\{y_i = 0, \forall(y_{ij}) = 0\}$ for negative slides, and $\{y_i = 1, \sum y_{ij} \geq 1\}$ when positive. This implies that instances in negative slides are *all* considered negative, with at least one positive instance for the rest. This assumption models the task as weakly supervised in nature with noisy labeled instances, making learning non-trivial. In the context of the few-shot setting, shot denotes the number of labeled slides per class *i.e.*, N-shot pairs of positive and negative slides for training (*e.g.*, 1,4,8,16 shots). During inference, the complete testing set is employed for weakly supervised classification.

Recent works have shown impressive results with vision-language (VL) models for few-shot learning. Here, we use the VL model PLIP [10], a fine-tuned CLIP [21] model trained with publicly curated digital pathology specific image-text pairs. CLIP-based frameworks comprise an image and text encoder based

on ResNet-50/ViT and Transformer for text feature modeling and are trained to align image and text features using a contrastive objective. The models also have an inherent ability for zero-shot inference using text descriptions or fixed prompts per category. Formally, let $\mathbf{z}^v = \mathbf{f}_\theta^v(x_i)$ be image features extracted by the visual encoder and $\{\mathbf{z}_i^t\}_{i=1}^C = \mathbf{f}_\phi^t(w_i)$ weight vectors from the text encoder, where C represents the number of classes derived from text descriptions $\{w_i\}$ *i.e.* "An H&E image of [CLASS]" with [CLASS] replaced with "*breast adenocarcinoma*", "*invasive ductal carcinoma*" or "*normal smooth mucosa*". The probability for each category can be obtained following:

$$q(y = i \mid x) = \frac{\exp\left(\mathbf{sim}(\mathbf{z}_i^t, \mathbf{z}^v)/\tau\right)}{\sum_{j=1}^K \exp\left(\mathbf{sim}(\mathbf{z}_j^t, \mathbf{z}^v)/\tau\right)}, \tag{1}$$

where τ is a learned temperature coefficient and $\mathbf{sim}(\cdot)$ the cosine similarity.

Prompt Initialization. To achieve improved generalization on downstream tasks with pre-trained VL-models without full fine-tuning, CoOp [30] alternatively learns continuous vectors appended to a given category prompts initialized as $\mathbf{p} = [P]_1, [P]_2, \ldots, [P]_K, [\text{CLASS}]$, with each $[P]_K \in \mathbb{R}^{d^t}$ a vector with same dimension (d^t) as word embeddings, and K the number of context tokens. Note that only $[P]$ is optimized using the cross-entropy loss given probabilities via Eq. (1). As pointed out by Lee *et al.* [15], learnable prompts may alter parts of the original model via the attention mechanism *i.e.*, all class and learnable embeddings will interact and affect the representation. Herein, we initialize the prompts as follows:

$$\mathbf{p} = [\mathbf{f}_\phi^{t(0)}; \mathbf{z}_i^t; \{\mathbf{p}_i\}_{i=1}^K], \tag{2}$$

where $\mathbf{f}_\phi^{t(0)} \in \mathbb{R}^{d^t}$ denotes the special token embedding [EOS] in the text encoder that acts as a feature aggregator. $\mathbf{z}_i^t \in \mathbb{R}^{N^t \times d^t}$ and $\mathbf{p}_i \in \mathbb{R}^{d^t}$ denote the text embeddings of length N^t tokens and the ith learnable prompts with K number of prompts per category, while d^t is the word embedding dimension. In contrast to CoOp which replaces token embeddings with learnable prompts, this mechanism encodes the entire text and concatenates the learnable prompts. Following [15], special tokens [EOS] used for K learnable prompts $\{\mathbf{p}_i\}_{i=1}^K$ are initialized as $p_i \sim \mathcal{N}(\mathbf{f}_\phi^{t(0)}, \sigma^2 I)$ with variance σ^2. This avoids constant initialization, with masked attention used during training to restrict attention flow from learnable prompts at different layers of the text encoder. We denote $\hat{\mathbf{z}}^t$ as the final projected text embedding of the K learned prompts *i.e.*, $\{\hat{\mathbf{z}}^t\}_{i=1}^K$.

Instance Feature Adaptation. In our framework, while learning optimized text-prompt features is crucial, we posit visual instance adaptation via $\mathcal{A}_\psi^v$ can further enhance performance alongside prompt-tuning, especially for few-shot classification following Gao *et al.* [8]. In contrast to full fine-tuning, $\mathcal{A}_\psi^v$ enables to avoid over-fitting on a few samples by adopting residual connections to dynamically combine existing pre-trained knowledge and that of the learnable prompt vectors. Formally, given visual embeddings $\mathbf{z}^v = \mathbf{f}_\theta^v(x_i)$, two layers of learnable linear transformations are integrated in $\mathcal{A}_\psi^v$ to transform $\mathbf{z}^v$ following:

$$\mathcal{A}_\psi^v(\mathbf{z}^v) = \text{ReLU}(\text{ReLU}((\mathbf{z}^v)^T \mathbf{W}_1)^T \mathbf{W}_2), \tag{3}$$

Algorithm 1 Proposed Method

Input: Pre-trained VL models $\{\mathbf{f}_\phi^t, \mathbf{f}_\theta^v\}$, visual adapter $\mathcal{A}_\psi^v$, dataset $\mathcal{D} = \{X_i, Y_i\}$

Initialize prompts $\mathbf{p}$. // Eq.(2)
for each batch in $\mathcal{D}_x$ **do**
  $\{\hat{\mathbf{z}}_i^t\}_{i=1}^K = \mathbf{f}_\phi^t(\mathbf{p})$ // prompt vectors
  $\{\mathbf{z}_i^v\}_{i=1}^N = \mathbf{f}_\theta^v(\mathbf{X_i})$ // visual slide vectors
  $\{\hat{\mathbf{z}}_i^v\}_{i=1}^N \longleftarrow \{\mathbf{z}_i^v\}_{i=1}^N$ // visual adaptation Eq.(4)
  $\hat{\mathbf{z}}_{\text{slide}}^v \longleftarrow [\{\hat{\mathbf{z}}_i^v\}_{i=1}^N, \{\hat{\mathbf{z}}_i^t\}_{i=1}^K]$ // pooling Eq.(6)
  $\hat{\mathbf{z}}_{\text{mean}}^t \longleftarrow \{\hat{\mathbf{z}}_i^t\}_{i=1}^K$ // prompt pooling Eq.(7)
  $q(y = i|X_i) \longleftarrow [\hat{\mathbf{z}}_{\text{mean}}^t, \hat{\mathbf{z}}_{\text{slide}}^v]$ // logits Eq.(8)
  // Update prompts and adapter
  $\Delta\mathbf{p}, \Delta\mathcal{A}_\psi^v \leftarrow \min \mathcal{L}_{\text{cross-entropy}}(q, Y)$
endfor

Output: Prompts $\{\mathbf{p}_i\}_{i=1}^K$, Adapter $\mathcal{A}_\psi^v$

$$\hat{\mathbf{z}}^v = \alpha\mathcal{A}_\psi^v(\mathbf{z}^v) + (1-\alpha)\mathbf{z}^v, \tag{4}$$

with $\hat{\mathbf{z}}^v$ denoting the new adapted visual embeddings weighted by a residual ratio α to adjust the degree of conserving knowledge in the original projected vectors, modeled as a learnable parameter (default $\alpha = 0.5$).

Context-Driven Instance Pooling. For this study, it is crucial to first aggregate the projected visual-instance features $\{\hat{\mathbf{z}}^v\}_{i=1}^n = \mathbf{f}_\theta^v(\{x_i\}_{i=1}^n)$. Herein, we propose to pool all instance vectors to a single representation for slide-level classification by weighting each instance with the learned K prompts. Formally, given $\hat{\mathbf{z}}_\mathbf{c}^\mathbf{t} \in \mathbb{R}^{K\times d^t}$ and $\hat{\mathbf{z}}^v \in \mathbb{R}^{N\times d^v}$, textual learned prompts per class c and instance features, where d^t & d^v are the same size, we pool following:

$$\text{sim}_c(\hat{\mathbf{z}}^v, \hat{\mathbf{z}}_j^t) = \text{softmax}(\varphi(\hat{\mathbf{z}}^v) \cdot \varphi(\hat{\mathbf{z}}_j^t)/\tau), \tag{5}$$

$$\hat{\mathbf{z}}_{\{\text{slide},c\}}^v = \frac{1}{K}\sum_{j=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}\text{sim}_c(\hat{\mathbf{z}}_i^v, \hat{\mathbf{z}}_j^t)\cdot\hat{\mathbf{z}}_i^v\right), \tag{6}$$

where τ is the temperature hyper-parameter of the pre-trained model and $\text{sim}_c(\cdot)$ denotes the cosine similarity between $l2$-normalized (via φ) slide instances and the k-th prompt embedding, including softmax normalization applied on N instances in $\hat{\mathbf{z}}^v$ to obtain scores per instance for each class c. The scores are then used to pool across visual embeddings with the final average embedding/class weighted by the number of K prompts. Our pooling strategy is motivated by the fact that rather than introducing an additional learnable visual pooling module (*e.g.*, AbMIL [11]), the learned prompts can be used to capture different instance features related to the text. Consequently, given the slide level embedding $\hat{\mathbf{z}}_{\{slide,c\}}^v$, we optimize all learnable parameters modules similar to Eq. (1) based on the mean context embedding of learned prompts. As opposed to using a fixed or separate prompt learner for slide-level classification, the mean of K prompts is used as follows:

$$\hat{\mathbf{z}}^t_{\text{mean}} = l_2^{\text{norm}}\left(\frac{1}{K}\sum_{i=1}^{K}\hat{\mathbf{z}}^t_i\right), \tag{7}$$

$$q(y = i \mid X) = \frac{\exp\left(\text{sim}(\hat{\mathbf{z}}^t_{\text{mean}}, \hat{\mathbf{z}}^v_{\text{slide}})/\tau\right)}{\sum_{j=1}^{C}\exp\left(\text{sim}(\hat{\mathbf{z}}^t_{\text{mean},j}, \hat{\mathbf{z}}^v_{\text{slide}})/\tau\right)}. \tag{8}$$

The proposed framework is summarized in Algorithm 1, which uses the final slide-level logits $q(\cdot)$ to optimize the prompts and adapter parameters.

3 Experiments

Datasets. We validate the effectiveness of the proposed approach on the publicly available NCT [13] and Camelyon16 [2] datasets. NCT dataset was proposed to classify human colorectal cancer (CRC) and normal tissue. It consists of $100,000/7,180$ image patches (224×224) in 9 tissue classes extracted from hematoxylin & eosin (H&E) stained Whole Slide Images (WSIs) from 86/50 patients for training/validation. To simulate the few-shot weak setting, we constructed positive/negative slides (bags) with a fixed bag size of ($N = 32$) and a positive instance ratio of 5%. Specifically, cancer-associated classes *cancer-associated stroma* and *colorectal adenocarcinoma epithelium* were used to sample positive instances for bag construction, with the rest used for negative bags. Camelyon16 (CM16) was introduced for breast cancer metastasis detection in lymph nodes. It comprises 400 H&E-stained WSIs of lymph nodes and is divided into 271 WSIs for training and 130 for testing. During pre-processing, each WSI was cropped into non-overlapping patches ($\times20$ magnification) resulting in a total of 3.2M patches. Each bag contains an average of 8,800 patches, with a maximum of 20,000 patches per bag. In contrast to NCT, no bag construction is necessary as WSIs have slide-level labels.

Implementation Details. We compare the proposed method against (i) Linear-Probing (MaxPool), (ii) Linear-Probing (AbMIL [11]), (iii) CoOp [30], (iv) CoOp-GCE [27], and (v) Adapter [8] with frozen VL models (ViT-B/32 PLIP [10]). All models employ prompt templates ("An H&E image of [CLASS]") based on class categories of each dataset *i.e.* "benign normal tissue" & "colorectal adenocarcinoma" on NCT, and "metastatic breast cancer" on Camelyon16. We also include AbMIL trained using all samples (100%) for efficiency, and prompt-tuning baselines employ AbMIL [11] pooling by default. The Adam optimizer was used with batch size 1 and a learning rate of 0.001 for all methods trained for 50 epochs across different N-shot settings. For a fair comparison, the number of prompts-pairs was set as $K = 4$ in all prompting baselines.

4 Main Results

Table 1 presents the results of our approach on two evaluated datasets. The proposed method achieves significant gains against the zero-shot baseline and

Table 1: Performance comparison on NCT and CM16. Slide-level classification based on average AUC repeated 3 times with different model initialization. The best and second best results are denoted in red and blue. † and †† denotes using AbMIL & MaxMIL pooling.

Dataset	NCT				Camelyon16			
Methods/N-shot	4	8	16	Avg	4	8	16	Avg
AbMIL (100%)	-	-	-	96.43	-	-	-	85.43
MaxMIL [25]	53.36	54.39	58.41	55.39	55.14	64.14	61.68	60.32
AbMIL [11]	52.17	48.26	66.17	55.53	63.68	66.05	78.91	69.55
PLIP [10]	-	-	-	56.31	-	-	-	65.31
+ LinearProbe [21] ††	59.05	57.08	58.00	58.04	54.86	63.97	57.19	58.67
+ LinearProbe [21] †	52.50	60.32	67.90	60.91	56.68	57.56	57.10	57.45
+ CoOp [30] †	69.58	81.63	84.41	78.54	64.12	67.47	71.25	67.28
+ CoOp-GCE [27] †	69.56	83.29	84.92	79.59	62.99	65.25	68.15	65.46
+ Adapter [8] †	70.89	77.82	80.40	76.03	59.98	68.72	75.54	68.08
+ **Ours**	87.51	86.99	89.94	88.48	64.12	79.40	80.78	74.43

Table 2: Instance-level performance. The best and second best results are denoted in red and blue.

	NCT			CM16		
Methods	4-shot	8-shot	16-shot	4-shot	8-shot	16-shot
AbMIL	58.33	68.91	56.06	71.83	84.88	87.44
PLIP						
+ CoOp	83.92	82.80	83.32	78.56	71.83	78.27
+ **Ours**	87.24	88.05	90.16	87.94	92.02	93.62

Table 3: Effect of modules on CM16 dataset. *C-Pooling* denotes context-driven pooling.

Methods (PLIP)	C-Pooling	Adaptation.	AUC
+ CoOp			71.25
+ Ours w/ Mean Pool		✓	55.34
+ Ours w/ AbMIL		✓	75.85
+ Ours	✓		67.53
+ Ours	✓	✓	80.78

outperforms the weakly supervised AbMIL in average performance, especially in the 16-shot setting on NCT. Compared to AbMIL(100%), performance was fairly comparable *i.e.*, −7.95%, suggesting that when trained on more samples, our prompt-based learning can further enhance performance. Notably, among the prompting baselines, CoOp-GCE reports the best overall performance, with Adapter and CoOp in other evaluated few-shot settings. Recall that all compared prompt-based baselines employ an instance pooling module (AbMIL), which enables improved learning. We observed that the exclusion of the default pooling mechanism results in significantly lower performance as shown for Linear-probing (MaxMIL). The benefit of the proposed context-driven pooling is shown despite not incurring extra learning parameters.

Furthermore, we report the best results on the challenging Camelyon16 and observed the difference in results between the zero-shot baseline and our method was less pronounced *i.e.*, 65% vs. 74% compared to results on NCT. Interestingly, Adapter and AbMIL baselines show clear gains over prompting; we attribute this to the larger number of instances on this dataset *i.e.*, mean bag size 8800, especially in the 16-shot setting. Nevertheless, we posit that text description initialization plays an important role in the efficient transfer of image-text in the few-shot setting, despite learning under the presence of tissues with uneven distribution across slides. To validate the robustness of learning with weak labels, we present instance-level results in Table 2. Zero-shot PLIP reports good performance without tuning, suggesting the model is already robust for instance

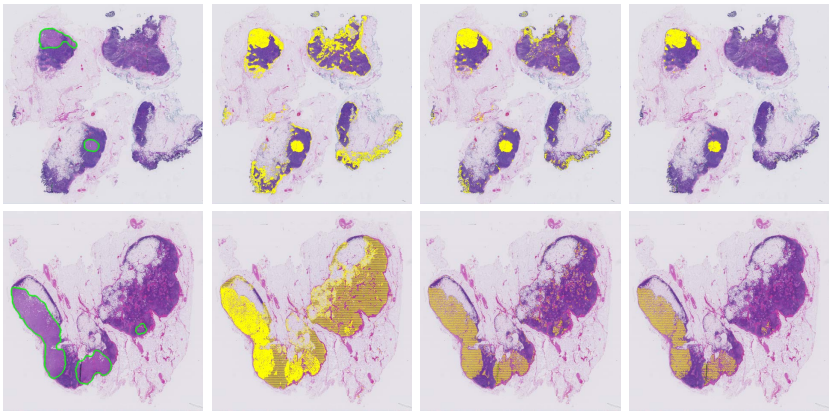

Fig. 2: Illustration of tumor ROI localization on Camelyon16 test samples. The green annotation denotes the ground-truth, and yellow are the compared model instance predictions (16-shot trained).

level tasks; but is not well-suited for slide-level multiple-instance inference as higlighted in Table 1. Furthermore, we analyzed instance-level performance per tissue type on NCT and report scores that are consistent with slide-level labels *i.e.*, lower scores were assigned to non-cancerous instances in positive slides compared to the baselines (Appendix Fig. S2).

To better highlight slide-level tumor localization, Figure 2 shows we can segment relevant tumor ROI even when trained with few-samples. Note that both AbMIL and CoOp exhibit a higher number of false positive predictions compared to ours in both exemplars. We also validate the robustness of the learned prompt vectors wherein a new query description (*e.g.*, Debris tissue, Mucus tissue, etc.) can be used to retrieve the top-scoring instances for the given query (Appendix Fig. S3). We intend to show that prompt-tuning in the MIL setting still retains the base performance of the VL model and can further enhance the performance as highlighted by more correctly retrieved samples.

Effect of Context-pooling and Adaptation. As shown in Table 3, compared with using the mean of all instances (MeanPool) as the slide feature, the proposed pooling strategy can boost the performance of the baselines. Further, we observed more gains over CoOp when using AbMIL with adaptation. Note that while using a single strategy shows improved scores, the combination of the proposed pooling method and visual adaptation was overall better and complementary.

5 Conclusion

In this paper, we introduced prompt-tuning for histology classification with learned prompt vectors for instance feature pooling in whole slide images. We

show that the combination of feature adaptation and context-based pooling can enhance learning in data-deficient settings, and is more efficient and comparable to standard fine-tuning with multiple instance methods. Further exploring the transferability of other pre-trained pathology VL models, subtyping tasks on large datasets, including prompting with different pre-trained image encoders is a topic of future research.

Acknowledgements. This work was supported by IITP grant funded by the Korean government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and (No.RS-2024-00439264, Development of High-Performance Machine Unlearning Technologies for Privacy Protection), Smart Health Care Program funded by the Korean National Police Agency (220222M01).
Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amores, J.: Multiple instance classification: Review, taxonomy and comparative study. Artificial intelligence **201**, 81–105 (2013)
2. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
3. 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: CVPR. pp. 16144–16155 (2022)
4. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV. pp. 104–120. Springer (2020)
5. Chikontwe, P., Nam, S.J., Go, H., Kim, M., Sung, H.J., Park, S.H.: Feature recalibration based multiple instance learning for whole slide image classification. In: MICCAI. pp. 420–430. Springer (2022)
6. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine Learning with Applications **7**, 100198 (2022)
7. Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Frontiers in medicine p. 264 (2019)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. IJCV pp. 1–15 (2023)
9. He, L., Long, L.R., Antani, S., Thoma, G.R.: Histology image analysis for carcinoma detection and grading. Computer methods and programs in biomedicine **107**(3), 538–556 (2012)
10. 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 pp. 1–10 (2023)
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136. PMLR (2018)

12. 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: ICML. pp. 4904–4916. PMLR (2021)
13. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. https://doi.org/10.5281/zenodo **1214456** (2018)
14. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: ICLR (2022)
15. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: CVPR. pp. 1401–1411 (2023)
16. Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J.: Dt-mil: Deformable transformer for multi-instance learning on histopathological image. In: MICCAI. pp. 206–216. Springer (2021)
17. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL. pp. 4582–4597 (2021)
18. 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: CVPR. pp. 19764–19775 (2023)
19. 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)
20. Qu, L., Fu, K., Wang, M., Song, Z., et al.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification (2024)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
22. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: EMNLP. pp. 4222–4235 (2020)
23. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67**, 101813 (2021)
24. Srinidhi, C.L., Martel, A.L.: Improving self-supervised learning with hardness-aware dynamic curriculum learning: an application to digital pathology. In: CVPR. pp. 562–571 (2021)
25. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)
26. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: CVPR. pp. 7959–7971 (2022)
27. Wu, C.E., Tian, Y., Yu, H., Wang, H., Morgado, P., Hu, Y.H., Yang, L.: Why is prompt tuning for vision-language models robust to noisy labels? In: CVPR. pp. 15488–15497 (2023)
28. Zhang, J., Kapse, S., Ma, K., Prasanna, P., Saltz, J., Vakalopoulou, M., Samaras, D.: Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. In: MICCAI. vol. 14227, pp. 624–634. Springer Nature Switzerland (2023)
29. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)