FALFormer: Feature-aware Landmarks self-attention for Whole-slide Image Classification

Doanh C. Bui, Trinh Thi Le Vuong and Jin Tae Kwak (✉)

School of Electrical Engineering, Korea University, Seoul, Republic of Korea
jkwak@korea.ac.kr

Abstract. Slide-level classification for whole-slide images (WSIs) has been widely recognized as a crucial problem in digital and computational pathology. Current approaches commonly consider WSIs as a bag of cropped patches and process them via multiple instance learning due to the large number of patches, which cannot fully explore the relationship among patches; in other words, the global information cannot be fully incorporated into decision making. Herein, we propose an efficient and effective slide-level classification model, named as FALFormer, that can process a WSI as a whole so as to fully exploit the relationship among the entire patches and to improve the classification performance. FALFormer is built based upon Transformers and self-attention mechanism. To lessen the computational burden of the original self-attention mechanism and to process the entire patches together in a WSI, FALFormer employs Nyström self-attention which approximates the computation by using a smaller number of tokens or landmarks. For effective learning, FALFormer introduces feature-aware landmarks to enhance the representation power of the landmarks and the quality of the approximation. We systematically evaluate the performance of FALFormer using two public datasets, including CAMELYON16 and TCGA-BRCA. The experimental results demonstrate that FALFormer achieves superior performance on both datasets, outperforming the state-of-the-art methods for the slide-level classification. This suggests that FALFormer can facilitate an accurate and precise analysis of WSIs, potentially leading to improved diagnosis and prognosis on WSIs.

Keywords: WSI classification · Nyström self-attention · Transformer

1 Introduction

In recent years, slide-level whole-slide image (WSI) classification has drawn considerable attention due to its crucial role in clinics for disease diagnosis and prognosis [10]. Given that WSIs are gigabytes in size, obtaining pixel-level annotations and conducting patch-level classification poses significant challenges to the field of computational pathology. The common strategy to handle and process WSIs is to adopt the multiple instance learning (MIL) paradigm, in which

WSIs are divided into a set of disjoint patches, representing WSIs as a bag of patches or instances, and the information from the patches are extracted, selected, and/or aggregated to produce the slide-level prediction. There are two main MIL-based approaches including instance-based [7,1] and bag embedding-based models [8,9,11,12,16]. Instance-based models conduct path-level predictions and then aggregate the results to produce the final prediction for a WSI, while bag embedding-based models map the patches in a bag into one embedding vector and make a prediction based on it. It has recently shown that instance-based models are less efficient than bag embedding-based models [13,12]. Bag embedding-based models are mostly built based upon an attention mechanism and Transformer architecture. For example, AB-MIL [8] learns to assign a weight for each patch using the attention mechanism, and then performs a weighted average to aggregate all the patch embeddings. Similarly, CLAM [11] conducts an auxiliary task that clusters the top most-attended patch embeddings as positive patches and the top least-attended patch embeddings as negative patches to constrain and refine the feature space. DTFD-MIL [16] is a two-stage MIL-based model, involving sub-MIL and global MIL models, that sought to handle overfitting problems due to the limited number of WSIs. The original patch bag undergoes random splitting to create multiple sub-bags, of which each is processed and aggregated using a sub-MIL model. The representative embeddings from the sub-bags are fed into a global MIL model. However, long-range dependencies among patches have not been exploited well, as bag embedding is produced by computing a weighted sum of all patches. To address this, some recent studies attempt to make use of self-attention. For instance, TransMIL [12] adopts an architecture of Transformer with positional encoding to retain the spatial information of the cropped patches by reshaping them into 2-D image space and applying multiple learnable convolutions, and utilizes vanilla Nyström Attention to approximate self-attention among patch features. HIGT [5] introduces a strategy of using clustering and pooling to reduce the number of patches and applies a variant of self-attention. MSPT [3] performs clustering to reduce the number of patches uses them as queries for self-attention.

Though successful, these previous MIL-based models, by and large, do not fully explore and utilize all the available patches due to the enormous number of the patches and the computational cost, which likely limits the capability of the model and the subsequent decision making. Hence, advanced methods that can efficiently and effectively process a WSI as a whole or the entire patches together in efficient manner are needed to improve the accuracy and efficiency of the slide-level image classification.

Herein, we propose a **F**eature-**A**ware **L**andmarks Trans**Former** (`FALFormer`) for efficient and effective slide-level image classification. FALFormer is built based upon Transformers and self-attention mechanisms. To reduce the computational burden and to process the entire patches in a WSI, FALFormer adopts Nyström self-attention [15] which approximates the computation by using a smaller number of tokens or landmarks. For effective learning, FALFormer introduces a **F**eature-**A**ware **L**andmarks Nyström **S**elf-**A**ttention (`FALSA`), which

enhances the representation power of the landmarks so as to better approximates the self-attention computation, leading to improved classification performance. Two public datasets, including CAMELYON16 and TCGA-BRCA, are employed to evaluate FALFormer. The experimental results demonstrate that FALFormer is able to conduct the slide-level image classification in an accurate and robust manner and outperforms the state-of-the-art models. Our implementation is available at [1].

2 Methodology

In this section, we present FALFormer for the WSI classification/sub-typing problem. Let $\mathbf{X}$ be a Giga-sized WSI and $\mathbf{Y}$ be the slide-level class label of $\mathbf{X}$. The objective of our study is to develop a Transformer-based model $\mathcal{T}(\cdot)$, i.e., FALFormer, which fully exploits entire patches tiled from the WSI with spatially-aware landmarks, and predict the slide-level label: $\mathbf{Y} = \mathcal{T}(\mathbf{X})$. The overview of FALFormer is illustrated in Fig. 1.

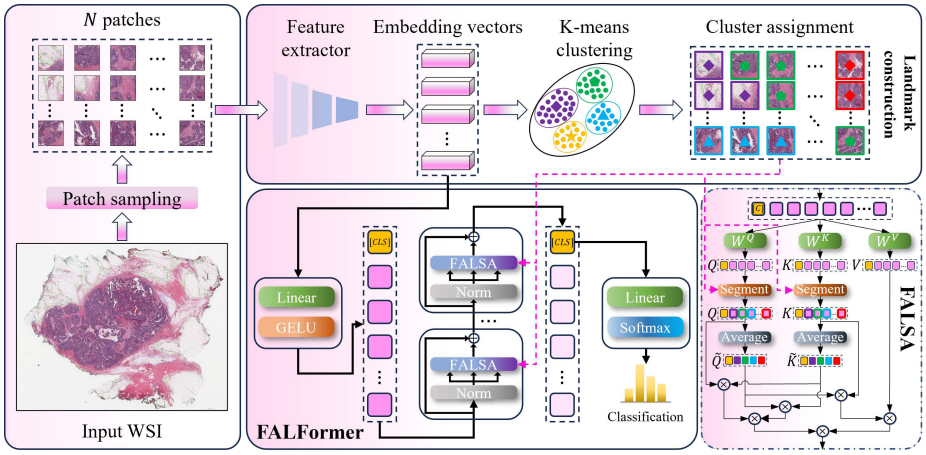


Fig. 1. Overview of FALFormer. A WSI is first segmented and tiled into patches based on foreground regions. Then, patch embeddings are extracted and divided into a number of segments by using K-means clustering. FALFormer computes representative landmarks using the segments and use them to process the entire patch embeddings in an efficient and effective mannder for the slide-level classification.

2.1 Feature-aware Landmarks Transformer

Given $\mathbf{X}$, we first remove background to retrieve tissue regions. Subsequently, the tissue regions are tiled into a bag of N patches $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ where $\mathbf{p}_i$ denotes the ith patch and N depends on each WSI. From each patch $\mathbf{p}_i$, an embedding

[1] https://github.com/caodoanh2001/FALFormer

vector $\mathbf{h}_i \in \mathbb{R}^{1\times d_f}$, a.k.a. a patch token, is produced using a feature extractor $\mathcal{F}(\cdot)$ where d_f is the dimension of the embedding vector and can vary depending on the choice of $\mathcal{F}(\cdot)$. Finally, a set of patch tokens $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^{N}$ is obtained.

FALFormer receives $\mathbf{H}$ and conducts a linear projection `FC` followed by a `GELU` activation to map the dimension of the patch tokens to the dimension of the model space d_{model}. The resultant patch tokens are concatenated with a learnable `[CLS]` token, denoted as $\mathbf{h}^{(0)}_{\texttt{[CLS]}} \in \mathbb{R}^{1\times d_{model}}$, producing $\mathbf{H}^{(0)} = \texttt{Concatenate}\big(\mathbf{h}^{(0)}_{\texttt{[CLS]}}, \texttt{GELU}(\texttt{FC}(\mathbf{H}))\big) \in \mathbb{R}^{(N+1)\times d_{model}}$. Then, $\mathbf{H}^{(0)}$ undergoes a stack of L Transformer layers given by: $\mathbf{H}^{(i)} = \texttt{Transformer}^{(i)}\big(\mathbf{H}^{(i-1)}\big), 1 \leq i \leq L$ where $\texttt{Transformer}^{(i)}(\cdot)$ consists of a normalization layer `LayerNorm` and a `FALSA`. `LayerNorm` learns affine transform parameters for the patch tokens: $\mathbf{H}' = \texttt{LayerNorm}\big(\mathbf{H}^{(i-1)}\big)$. `FALSA` is used to approximate the self-attention computation for $\mathbf{H}'$ with a skip connection: $\mathbf{H}^{(i)} = \mathbf{H}^{(i-1)} + \texttt{FALSA}(\mathbf{H}')$. Finally, only the `[CLS]` token is utilized for the prediction: $\mathbf{Y} = \texttt{Classifier}\Big(\texttt{LayerNorm}\big(\mathbf{h}^{(L)}_{\texttt{[CLS]}}\big)\Big)$.

2.2 Revisiting Nyström self-attention

Nyström self-attention [15] is an approach to approximate the self-attention computation. Given a sequence of patch tokens $\mathbf{H}$ and three learnable projection matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$, the standard self-attention computes the query $\mathbf{Q} \in \mathbb{R}^{N\times d_q}$, key $\mathbf{K} \in \mathbb{R}^{N\times d_k}$, and value $\mathbf{V} \in \mathbb{R}^{N\times d_v}$ as follows: $\mathbf{Q} = \mathbf{H}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{H}\mathbf{W}^K$, $\mathbf{V} = \mathbf{H}\mathbf{W}^V$ where N is the number of tokens (N is large) and d_q, d_k, and d_v is the dimension of the query, key, and value vector, respectively. This requires to compute attention weights $att = \texttt{softmax}\Big(\frac{QK^T}{\sqrt{d_k}}\Big)$, which may cause the out of memory problem due to the large N. To lessen the computational burden, Nyström self-attention reduces N to N' ($N' \ll N$) for $\mathbf{Q}$ and $\mathbf{K}$ by grouping N tokens into N' segments where each segment contains $N_k = N/N'$ tokens, producing sets of segments $\{Q^{(i)}\}_{i=1}^{N'}$ and $\{K^{(i)}\}_{i=1}^{N'}$. The tokens are averaged within each segment, forming landmarks $\tilde{\mathbf{Q}} = \big\{\frac{1}{N_k}\Sigma Q^{(i)}\big\}_{i=1}^{N'}$ and $\tilde{\mathbf{K}} = \big\{\frac{1}{N_k}\Sigma K^{(i)}\big\}_{i=1}^{N'}$. Then, Nyström self-attention can be formulated as:

$$\begin{gathered}\tilde{F} = \texttt{softmax}\Big(\frac{\mathbf{Q}\tilde{\mathbf{K}}^T}{\sqrt{d_q}}\Big), \tilde{A} = \texttt{softmax}\Big(\frac{\tilde{\mathbf{Q}}\tilde{\mathbf{K}}^T}{\sqrt{d_q}}\Big)^+, \tilde{B} = \texttt{softmax}\Big(\frac{\tilde{\mathbf{Q}}\mathbf{K}^T}{\sqrt{d_q}}\Big), \\ \tilde{F} \in \mathbb{R}^{N\times N'}, \tilde{A} \in \mathbb{R}^{N'\times N'}, \tilde{B} \in \mathbb{R}^{N'\times N}, \\ \mathbf{H} = (\tilde{F} \times \tilde{A}) \times (\tilde{B} \times \mathbf{V}),\end{gathered} \tag{1}$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudoinverse function. The computational complexity of the standard self-attention computation is $\mathcal{O}(N^2)$ because the scale-dot matrix multiplication should be done for all N tokens, which is inappropriate in the case of a large N. Nyström self-attention (Eq. 1) has the computational complexity of $\mathcal{O}(N)$, which is much smaller than $\mathcal{O}(N^2)$ if $N' \ll N$. Though successful, there still remains a question of **how to choose the landmarks to achieve better performance in the context of WSI?** Nyström

self-attention groups the N tokens into N' segments in order from top to bottom. We hypothesize that the better landmarks we choose, the better approximation we obtain, leading to improved classification performance.

2.3 Feature-aware Landmarks Nyström Self-Attention (FALSA)

FALSA chooses the best representative landmarks as follows. First, it defines the maximum number of segments, denoted as N_s ($N_s \ll N$). Second, it utilizes the K-means clustering algorithm to find N_s centroids and to divide the patch tokens $\mathbf{H}$ into N_s segments: $\mathbf{S} = \texttt{K-means}(\mathbf{H}, N_s)$ where $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{N}, s_i \in \{1, 2, ..., N_s\}$, denotes the set of segment IDs for $\mathbf{H}$. Third, tokens in the query $\mathbf{Q}$ and the key $\mathbf{K}$ are segmented based on their segment IDs $\mathbf{S}$, except for the [CLS] tokens, i.e., $\mathbf{q}_{\texttt{[CLS]}}$ and $\mathbf{k}_{\texttt{[CLS]}}$, to form $\mathcal{Q} = \{Q^{(j)}\}_{j=1}^{N_s}$, $\mathcal{K} = \{K^{(j)}\}_{j=1}^{N_s}$. In which, $Q^{(j)}$ and $K^{(j)}$ denote sets of query and key tokens, respectively, belonging to the same j^{th} segment: $Q^{(j)} = \{\mathbf{q}_i \in \mathbf{Q} | \mathbf{s}_i = j\}_{i=1}^{N}$, $K^{(j)} = \{\mathbf{k}_i \in \mathbf{K} | \mathbf{s}_i = j\}_{i=1}^{N}$. Fourth, it computes N_s landmarks for the query ($\tilde{\mathcal{Q}}$) and key ($\tilde{\mathcal{K}}$) by computing the average of the tokens within each segment: $\tilde{\mathcal{Q}} = \left\{\frac{1}{C(j)} \sum Q^{(j)}\right\}_{j=1}^{N_s}$, $\tilde{\mathcal{K}} = \left\{\frac{1}{C(j)} \sum K^{(j)}\right\}_{j=1}^{N_s}$ where $C(j)$ denotes the the number of the patch tokens belonging to the j^{th} segment. Fifth, it concatenates $\tilde{\mathcal{Q}}$ and $\tilde{\mathcal{K}}$ with [CLS] tokens, forming $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{K}}$: $\tilde{\mathbf{Q}} = \texttt{Concatenate}(\mathbf{q}_{\texttt{[CLS]}}, \tilde{\mathcal{Q}}), \tilde{\mathbf{K}} = \texttt{Concatenate}(\mathbf{k}_{\texttt{[CLS]}}, \tilde{\mathcal{K}})$. Last, $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{K}}$ are utilized to compute the Nyström self-attention as described in Eq. 1.

3 Experiments and Results

3.1 Datasets

CAMELYON16 [4]. The dataset was obtained from the CAMELYON16 challenge, which was designed to evaluate algorithms for metastasis detection. There are 399 WSIs in total, and the official train-test split is used. Specifically, 216, 54, and 129 WSIs are employed for training, validation, and testing, respectively. WSIs are tiled to 3,617,584 patches, with 9066.6±6273.6 patches per WSI. In this study, we use CAMELYON16 for tumor vs. non-tumor classification.

TCGA-BRCA. The dataset includes a total of 875 WSIs for breast cancer sub-typing, such as Invasive Ductal Carcinoma (IDC) versus Invasive Lobular Carcinoma (ILC). These annotated WSIs are available on the NIH Genomic Data Commons Data Portal. Following [2], we use the ratio of 0.8:0.1:0.1 for the train-val-test split, which are 715 and 79 WSIs for training and validation, respectively, and 81 WSIs for testing. 2,672,891 patches are generated from the WSIs, with 2567.6±1592.8 patches per WSI.

3.2 Implementation Details

To obtain bag of patches from a WSI, we follow the pipeline provided in the previous study [11]. For FALFormer, we set the number of transformer layers L to 2 and the model dimension d_{model} to 768, which are inspired by designs of Vision Transformers and TransMIL. The number of segments N_s is set to 256. For feature extraction, we utilize two pre-trained models: ResNet-50 [6], pre-trained on ImageNet1K, and CTransPath [14], which is a SwinT-based architecture pre-trained on histopathology datasets. ResNet-50 and CTransPath produce feature vectors of sizes $d_f = 1024$ and $d_f = 768$, respectively. FALFormer is trained for 20 epochs. During training, the RAdam optimizer is utilized, cross-entropy loss is adopted, and the EarlyStopping strategy is employed to halt training if the validation loss does not improve after 10 epochs. The best model is chosen based on a validation set. The experiment is conducted only once for FALFormer and other models under identical conditions, using the same random seed and environment.

3.3 Comparative Study

For comparison, we include two established slide-level classification models: CLAM [11] and TransMIL [12]. Both are MIL-based models. CLAM clusters positive and negative patch embeddings within a WSI based on attention scores to improve the bag representation. CLAM can contain a single attention branch (CLAM-SB) and multi-attention branches (CLAM-MB). TransMIL employs a stack of Transformer layers and a positional encoding to capture the spatial information among patch embeddings. We build these three models using the same patch sampling procedure and maintain the hyperparameters from the original works for a fair comparison with FALFormer.

3.4 Result and Discussions

We assessed the performance of FALFormer and thee MIL-based models on the two datasets (CAMELYON16 and TCGA-BRCA) using five evaluation metrics including Accuracy (Acc), F1 score (F1), Area under the ROC curve (AUC), Recall, and Precision. Table 1 demonstrates the tumor vs. non-tumor classification results on CAMELYON16. Overall, FALFormer with CTransPath achieved the best classification performance of 96.12% Acc, 0.958 F1, 0.983 AUC, 0.957 Recall, and 0.960 Precision, substantially outperforming other MIL-based models such as $\geq$2.32 Acc%, $\geq$0.023 F1, $\geq$0.005 AUC, $\geq$0.019 Recall, and $\geq$0.018 Precision. It is worth noting that, using ResNet50 as the feature extractor, there was a consistent performance drop for all the models under consideration. Nonetheless, FALFormer with ResNet50 was superior or comparable to other MIL-based models with ResNet50. These results indicate that the quality of WSI analysis may be dependent on the choice of the feature extractor and the superior performance of FALFormer is not due to a specific choice of the feature extractor.

Table 1. Results on CAMELYON16 dataset.

Encoder	Method	Acc (%)	F1	AUC	Recall	Precision
ResNet50	CLAM-SB [11]	**86.05**	**0.849**	0.910	**0.844**	0.857
	CLAM-MB [11]	82.95	0.806	0.813	0.791	0.847
	TransMIL [12]	82.20	0.808	0.869	0.805	0.813
	FALFormer (ours)	**86.05**	0.848	**0.934**	0.840	**0.860**
CTransPath	CLAM-SB [11]	88.37	0.875	0.935	0.870	0.881
	CLAM-MB [11]	93.80	0.935	0.968	0.938	0.931
	TransMIL [12]	93.80	0.933	0.978	0.926	0.942
	FALFormer (ours)	**96.12**	**0.958**	**0.983**	**0.957**	**0.960**

Table 2 shows the breast cancer sub-typing results on TCGA-BRCA. Similar to the results on CAMELYON16, FALFormer outperformed the three competitors regardless of the choice of the feature extractor, highlighting the strength of FALFormer. FALFormer with CTransPath, in particular, obtained the best classification performance of 96.30% Acc, 0.937 F1, 0.970 AUC, 0.906 Recall, and 0.978 Precision. In a head-to-head comparison between ResNet50 and CTransPath, CTransPath always gave a substantial performance gain for FALFormer and other models except Recall for TransMIL; for instance $\geq$3.70% Acc, $\geq$0.060 F1, $\geq$0.017 AUC, and $\geq$0.059 Precision.

Table 2. Results on TCGA-BRCA dataset.

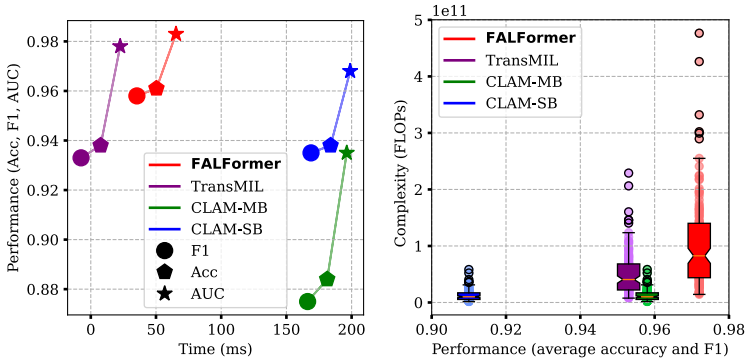
Encoder	Method	Acc (%)	F1	AUC	Recall	Precision
ResNet50	CLAM-SB [11]	90.12	0.817	0.926	0.773	0.900
	CLAM-MB [11]	91.36	0.844	0.942	0.805	**0.912**
	TransMIL [12]	88.89	0.787	0.932	**0.888**	0.742
	FALFormer (ours)	**92.59**	**0.877**	**0.945**	0.860	0.899
CTransPath	CLAM-SB [11]	95.06	0.914	0.958	0.875	0.971
	CLAM-MB [11]	95.06	0.914	**0.970**	0.875	0.971
	TransMIL [12]	93.83	0.888	0.949	0.844	0.964
	FALFormer (ours)	**96.30**	**0.937**	**0.970**	**0.906**	**0.978**

Moreover, we conducted ablation experiments to evaluate the effectiveness of `FALSA`. On CAMELYON16 and TCGA-BRCA, the performance of FALFormer with and without FALSA was measured. FALFormer without FALSA utilizes the original Nyström self-attention as described in Section 2.2. The results of the ablation experiments are presented in Table 3. It is obvious that the adoption of `FALSA` consistently enhances the classification performance regardless of the dataset and the feature extractor except for Acc, Precision on CAMELYON16 (ResNet50), and AUC, Recall on TCGA-BRCA (CTransPath). This indicates that the quality of landmarks has a direct bearing on the quality of the approximation of the self-attention and the final classification performance. Equipped with stronger landmarks by `FALSA`, FALFormer can facilitate improved analysis of WSIs.

Table 3. Ablation results demonstrating the effectiveness of `FALSA`.

Encoder	Self-attention	CAMELYON16					TCGA-BRCA				
		Acc (%)	F1	AUC	Recall	Precision	Acc (%)	F1	AUC	Recall	Precision
ResNet50	Nyström	86.05	0.824	0.905	0.824	0.886	90.12	0.827	0.922	0.797	0.873
	`FALSA`	86.05	0.848	0.934	0.840	0.860	92.59	0.877	0.945	0.860	0.899
CTransPath	Nyström	94.57	0.941	0.948	0.933	0.953	95.06	0.926	0.971	0.946	0.909
	`FALSA`	96.12	0.958	0.983	0.957	0.960	96.30	0.937	0.970	0.906	0.978

We also compared the performance and complexity trade-off within FALFormer and other competitors. We calculated the processing time for a WSI with the highest number of cropped patches, GFLOPs for all WSIs, and VRAM usage on CAMELYON16 using CTransPath as the feature extractor. These metrics are depicted in Figure 2. FALFormer exhibits the highest GFLOPs and VRAM usage, possibly due to the usage of the entire patches. CLAM models had the smallest GFLOPs and VRAM usage but required the longest processing time. TransMIL demonstrated the shortest processing time and second largest GFLOPs and VRAM usage. The performance of TransMIL was, in general, inconsistent (Table 1 and 2). Hence, the complexity analysis reveals that FALFormer strikes an acceptable balance between efficiency and accuracy. All measurements are calculated using a single NVIDIA RTX A6000 GPU.



Method	GFLOPs	VRAM (M)
CLAM-SB	12.74 ± 9.92	2,568
CLAM-MB	12.75 ± 9.93	2,568
TransMIL	50.93 ± 39.02	10,010
FALFormer	104.09 ± 81.06	10,260

c)

Fig. 2. Comparison of Complexity-Performance Trade-off: (a) Performance (Acc, F1, AUC) versus processing time for the most complicated WSI, (b) FLOPs for processing all WSIs versus average Acc and F1, and (c) GFLOPs and VRAM usage.

4 Conclusion

This study introduces FALFormer, a Transformer-based model for efficient and effective WSI classification. FALFormer revisited the Nyström-based self-attention mechanism and proposed `FALSA`, which leverages the high-level patch features and K-means algorithm to enhance the representative power of the landmarks and the quality of the approximation of the self-attention computation. Equipped with FALSA, FALFormer demonstrates its effectiveness in analyzing WSIs and conducting the slide-level classification.

Acknowledgments

This study was supported by a grant of the National Research Foundation of Kroea (NRF) (No. 2021R1A2C2014557 and No. RS-2024-00397293).

Disclosure of Interests

The authors declare that they have no conflict of interest.

References

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
2. 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)
3. Ding, S., Wang, J., Li, J., Shi, J.: Multi-scale prototypical transformer for whole slide image classification. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 602–611. Springer Nature Switzerland, Cham (2023)
4. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., , the CAMELYON16 Consortium: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA **318**(22), 2199–2210 (12 2017). https://doi.org/10.1001/jama.2017.14585, `https://doi.org/10.1001/jama.2017.14585`
5. Guo, Z., Zhao, W., Wang, S., Yu, L.: Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 755–764. Springer Nature Switzerland, Cham (2023)

6. 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)
7. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. 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)
10. Li, X., Li, C., Rahaman, M.M., Sun, H., Li, X., Wu, J., Yao, Y., Grzegorzek, M.: A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. Artificial Intelligence Review **55**(6), 4809–4878 (2022)
11. 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)
12. 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)
13. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)
14. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
15. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14138–14148 (2021)
16. 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)