CEMIL: Contextual attention based efficient weakly supervised approach for histopathology image classification

Tawsifur Rahman
Johns Hopkins University
arahma34@jhu.edu

Alexander S. Baras
Johns Hopkins University School of Medicine
baras@jhmi.edu

Rama Chellappa
Johns Hopkins University
rchella4@jhu.edu

Whole slide image preprocessing

Whole slide image (WSI) preprocessing begins with automated segmentation of tissue regions. Each WSI is read into memory at a downsampled resolution, such as 20×, and converted from RGB to HSV colorspace. A binary mask for the tissue regions (foreground) is created by thresholding the saturation channel of the image after applying median blurring to smooth the edges. This mask is then refined with morphological closing to fill small gaps and holes. The approximate contours of the detected foreground objects are filtered based on an area threshold and stored for further processing, while the segmentation mask for each slide is available for optional visual inspection. Additionally, a human-readable text file is generated, listing the processed files and editable fields for key segmentation parameters, allowing manual adjustments if needed. After segmentation, the algorithm crops 256×256 patches from within the segmented contours at the specified magnification and stores them, along with their coordinates and slide metadata, in the hdf5 hierarchical data format. The number of patches extracted per slide varies significantly, ranging from hundreds for biopsy slides at 20× magnification to hundreds of thousands for large resection slides at 40× magnification.

Ablation study

Comparison of different k values

In our study, we also evaluated the performance across different k values using standard CEMIL, as detailed in **Table 1**. The Instructor models trained with different k values showed varying levels of accuracy and AUC across all datasets—TCGA-NSCLC, TCGA-BRCA, TCGA-RCC, and PANDA. For instance, at $k = 0.4$, the Instructor (Last-k) model achieved accuracies of 44.75%, 48.55%, 42.68%, and 66.18% and AUCs of 50.34%, 52.24%, 49.11%, and 71.06% for TCGA-NSCLC, TCGA-BRCA, TCGA-RCC, and PANDA, respectively. Comparatively, CEMIL consistently outperformed the Instructor models across all datasets and k values. Notably, at $k = 0.6$, CEMIL (Serial) achieved accuracies and AUCs of 91.23% and 95.02% for TCGA-NSCLC, 89.78% and 95.76% for TCGA-BRCA, 92.15% and 95.25% for TCGA-RCC, and 87.68% and 91.79% for PANDA, respectively, demonstrating its superior performance. These results underscore the robustness and effectiveness of the CEMIL model in enhancing both accuracy and AUC metrics across diverse histopathology datasets.

Overall, our findings demonstrate the critical role of incorporating multiple loss functions in the training process, which significantly boosts the performance of CEMIL models. This comprehensive approach aligns patch representations with class predictions more effectively, thus providing a substantial improvement over traditional methods.

Comparisons of different loss with Standard CEMIL

In our study, we also conducted ablation experiments to assess various loss combinations in training learner networks for standard CEMIL, as depicted in **Table 2**. The initial phase of the learner network training involved minimizing $\mathcal{L}_{\text{PR}}$, which aligns the hidden patch representations with those of the instructor network. Subsequent fine-tuning focused on minimizing $\mathcal{L}_{\text{CE}}$ for the classification task. Additionally, we examined joint training that simultaneously minimized both $\mathcal{L}_{\text{PR}}$ and $\mathcal{L}_{\text{CE}}$ to synchronize patch representations and class predictions. The most comprehensive strategy involved minimizing all three losses: $\mathcal{L}_{\text{PR}}$, $\mathcal{L}_{\text{PP}}$, and $\mathcal{L}_{\text{CE}}$. Our results indicate that the serial training approach using the combined losses $\mathcal{L}_{\text{PR}} + \mathcal{L}_{\text{PP}} + \mathcal{L}_{\text{CE}}$ yielded the highest performance. Specifically, this method achieved slightly reduced accuracies and AUCs than Gated-CEMIL of 92.50% and 96.87% for TCGA-NSCLC, 91.62% and 96.41% for TCGA-BRCA, 93.50% and 96.89% for TCGA-RCC, and 88.97% and 92.73% for PANDA, respectively. These findings underscore the consistent enhancement of model accuracy and AUC through comprehensive

Table 1. The performance of CEMIL and baseline with different k values, and k is the percentage of total number of patches (N) in each bag.

k values	Models	TCGA-NSCLC		TCGA-BRCA		TCGA-RCC		PANDA		CAMEYLON16	
		Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
k =0.4	ABMIL (Random-k)	60.00	65.35	66.00	72.56	63.46	68.11	61.68	66.75	63.53	68.72
	Instructor (Random-k)	62.02	67.35	68.18	74.56	65.46	70.11	63.68	68.75	65.53	70.72
	Instructor (First-k)	42.66	48.08	55.02	62.88	50.35	57.01	63.27	68.18	44.97	50.56
	Instructor (Last-k)	44.75	50.34	48.55	52.24	42.68	49.11	66.18	71.06	46.68	52.78
	CEMIL (Parallel)	69.77	74.33	76.95	81.12	75.24	79.55	75.07	81.01	72.84	78.55
	CEMIL (Serial)	75.55	80.25	79.46	83.05	81.33	85.08	75.98	79.44	77.33	82.11
k =0.5	ABMIL (Random-k)	65.34	68.14	67.15	72.09	64.46	69.24	68.84	73.38	68.12	73.18
	Instructor (Random-k)	67.34	70.14	69.15	74.09	66.46	71.24	70.84	75.38	70.12	75.18
	Instructor (First-k)	59.78	63.25	61.62	66.98	53.55	56.75	68.71	72.75	62.34	67.52
	Instructor (Last-k)	55.53	61.06	62.35	66.36	59.77	64.01	71.62	77.58	57.56	63.42
	CEMIL (Parallel)	85.44	88.23	83.15	86.18	82.46	85.33	80.51	84.04	88.67	92.71
	CEMIL (Serial)	89.01	92.41	88.01	91.57	86.68	90.05	81.42	85.19	93.65	96.88
k =0.6	ABMIL (Random-k)	80.55	83.23	81.77	84.78	80.57	85.08	75.68	79.18	83.12	87.23
	Instructor (Random-k)	82.55	85.23	83.77	86.78	82.57	87.08	77.68	81.18	85.12	89.23
	Instructor (First-k)	62.24	65.22	64.13	68.46	55.41	60.09	75.55	79.33	64.47	69.32
	Instructor (Last-k)	63.35	66.79	67.02	71.16	65.22	70.24	78.46	82.66	65.72	71.18
	CEMIL (Parallel)	89.55	94.15	87.01	92.13	90.35	94.69	85.10	91.22	92.55	95.83
	CEMIL (Serial)	91.23	95.02	89.78	95.76	92.15	95.25	87.68	91.79	97.02	99.25
k =0.8	ABMIL (Random-k)	83.66	86.05	81.35	84.68	83.02	88.33	77.43	80.71	86.24	89.54
	Instructor (Random-k)	85.66	88.05	83.35	86.68	85.02	90.33	79.43	82.71	88.24	91.54
	Instructor (First-k)	77.55	80.38	75.48	81.13	78.75	81.59	77.32	82.72	80.44	84.85
	Instructor (Last-k)	83.35	86.21	82.35	85.69	84.55	88.96	80.21	84.28	86.41	90.12
	CEMIL (Parallel)	88.25	91.46	86.04	89.58	89.08	93.98	87.11	90.24	91.35	95.02
	CEMIL (Serial)	90.35	92.56	88.13	92.32	91.24	93.78	87.01	89.16	95.35	97.68

loss minimization, demonstrating its robustness across diverse histopathology datasets.

Comparisons of different loss weight

In our study, we evaluated the performance of the proposed Instructor-Learner models using different loss variants and varying loss weights λ_1 and λ_2, as shown in **Table 3**. The table reveals that the Serial training approach consistently outperformed the Parallel approach across all datasets and loss weight combinations. For instance, Serial training with $\lambda_1 = 0.4$ and $\lambda_2 = 0.3$ achieved the highest performance with accuracies and AUCs of 92.50% and 96.87% for TCGA-NSCLC, 91.62% and 96.41% for TCGA-BRCA, 93.50% and 96.89% for TCGA-RCC, and 88.97% and 92.73% for PANDA. These results highlight the importance of choosing appropriate loss weight combinations to optimize model performance. Notably, the Serial approach with $\lambda_1 = 0.4$ and $\lambda_2 = 0.3$ outperformed other combinations, suggesting that balanced weighting of loss functions can significantly enhance both accuracy and AUC. The findings underscore the effectiveness of the Serial training method in achieving higher performance, emphasizing the need for strategic selection of loss weights in the training process to align patch representations and class predictions more effectively, thereby improving overall model performance.

Table 2. Performance comparison of proposed Instructor-Learner models using different Loss variants.

Training	Loss	TCGA-NSCLC		TCGA-BRCA		TCGA-RCC		PANDA	
		Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Parallel	$\mathcal{L}_{PR} + \mathcal{L}_{CE}$	88.43	93.68	85.23	90.77	89.45	93.15	84.22	91.25
Parallel	$\mathcal{L}_{PP} + \mathcal{L}_{CE}$	89.56	94.80	87.03	92.57	89.94	94.64	84.84	91.53
Parallel	$\mathcal{L}_{PR} + \mathcal{L}_{PP} + \mathcal{L}_{CE}$	91.00	95.35	88.76	93.23	91.12	95.43	85.60	92.75
Serial	$\mathcal{L}_{PR} + \mathcal{L}_{CE}$	90.89	95.78	89.23	94.11	91.54	95.28	84.23	90.52
Serial	$\mathcal{L}_{PP} + \mathcal{L}_{CE}$	91.22	96.10	90.05	94.93	91.72	95.46	87.90	94.20
Serial	$\mathcal{L}_{PR} + \mathcal{L}_{PP} + \mathcal{L}_{CE}$	**92.50**	**96.87**	**91.62**	**96.41**	**93.50**	**96.89**	**88.97**	**92.73**

Table 3. Performance comparison of proposed Instructor-Learner models using different Loss variants with varying loss weights λ_1 and λ_2.

Training	Loss weight	TCGA-NSCLC		TCGA-BRCA		TCGA-RCC		PANDA	
		Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Parallel	$\lambda_1 = 0.6, \lambda_2 = 0.3$	87.54	92.68	84.32	89.76	88.45	92.15	83.42	90.15
Parallel	$\lambda_1 = 0.4, \lambda_2 = 0.2$	88.67	93.80	86.12	91.57	88.94	93.64	83.04	90.43
Parallel	$\lambda_1 = 0.3, \lambda_2 = 0.4$	90.10	94.35	87.86	92.23	90.12	94.43	84.60	91.75
Serial	$\lambda_1 = 0.5, \lambda_2 = 0.3$	89.89	94.78	88.23	93.11	90.54	94.28	83.23	89.52
Serial	$\lambda_1 = 0.2, \lambda_2 = 0.6$	90.22	95.10	89.05	93.93	90.72	94.46	86.90	93.20
Serial	$\lambda_1 = 0.4, \lambda_2 = 0.3$	**92.50**	**96.87**	**91.62**	**96.41**	**93.50**	**96.89**	**88.97**	**92.73**