Key Patches Are All You Need: A Multiple Instance Learning Framework For Robust Medical Diagnosis

Supplementary Material

1. Additional Results

1.1. Top-k Search In The MIL Framework

We conducted experiments with three different k values in the *instance-level* approach: approximately 12.5%, 25%, and 50%. In the context of our experiments, which involved input images of 224×224 resolution and patches of 16×16 resolution, we dealt with a total of $N = 196$ patches. This implies that for the $k \approx 12.5\%$ configuration, we considered 25 patches; for $k = 25\%$, we worked with 49 patches, and for $k = 50\%$, we used 98 patches. To determine the optimal k value for the *top-k average* operator in the *instance-level* approach, we conducted experiments using various backbones on the validation set of the ISIC 2019 dataset. A summary of these experiments is shown in Figure 1.

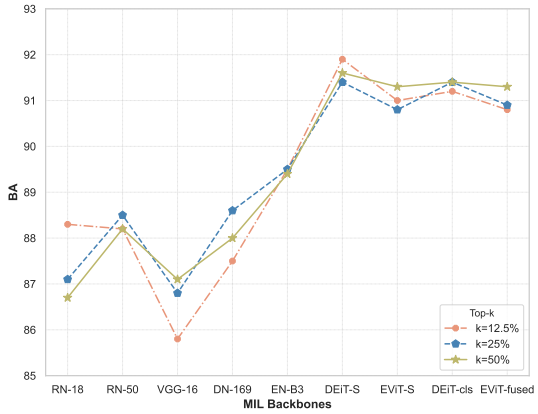

Figure 1. Search for the optimal k hyperparameter in the *instance-level top-k average* MIL pooling operator. We explored three values for the hyperparameter: $k \approx 12.5\%$, $k = 25\%$, and $k = 50\%$. Our experiments were conducted and evaluated on the validation set of the ISIC 2019 dataset, employing different MIL backbones. The backbones included RN-18, RN-50, VGG16, DN-169, EN-B3, DEiT-S, DEiT-cls (DEiT with the CLS token), EViT-S, and EViT-fused (EViT with the fused embedding). Notably, with EViT backbones, $k \approx 12.5\%$ resulted in only 9 patches, $k = 25\%$ retained 17 patches, and $k = 50\%$ maintained 34 patches. These results indicate that using more patches in the bag evaluation does not necessarily lead to better performance.

The plot in Figure 1 shows that the choice of k for the *top-k average* operator in the *instance-level* approach does not significantly impact the performance of the different MIL models. This observation suggests that not all patches within a dermoscopy image contribute equally to the classification task, indicating that the discriminative information lies within a (small) subset of image patches. Interestingly, the $k \approx 12.5\%$ and $k = 25\%$ scenarios consistently yield the highest BA results across different MIL backbones. Based on these results, we selected $k = 25\%$ as the default configuration for the *top-k average* pooling operator. To ensure a fair comparison between the *embedding-level* and *instance-level* approaches, we have also adopted $k = 25\%$ as the preferred setting for the *column-wise global top-k average* operator in the *embedding-level* approach.

1.2. EViT Experimental Setup Complementary Material

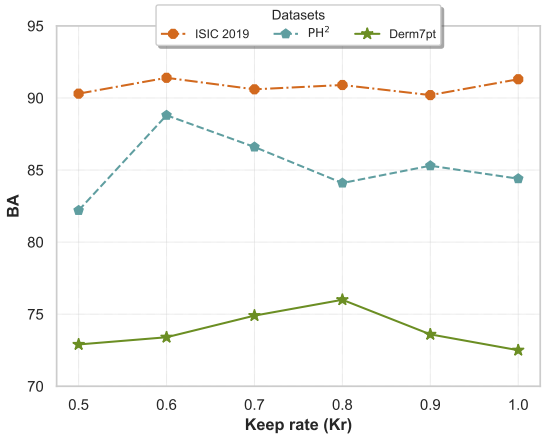

Figure 2. Performance comparison between EViT-S models with different keep rates. The experiments were conducted on the ISIC 2019 validation dataset, as well as on the two test datasets: PH^2 and Derm7pt, for the binary classification task of melanoma versus nevus. The x-axis represents the different K_r values, while the y-axis represents the corresponding BA results.

The combination of layers in which token reorganization takes place and the choice of K_r (keep rate) are the most critical hyperparameters in the EViT architecture. Since we decided to fix the token reorganization block in the 3^{rd}, 6^{th}, and 9^{th} layers, an extensive search for the best K_r configurations in the EViT-S model was required. Since the K_r hyperparameter plays a crucial role within the EViT architecture, we conducted a series of experiments to examine its impact on the EViT-S configuration. Figure 2 shows the BA

results across different K_r values for the ISIC 2019 validation set and both test datasets: PH^2 and Derm7pt. These results indicate that a higher K_r does not necessarily lead to better performance, especially on the test datasets. It is clear that K_r values of 0.6, 0.7, and 0.8 outperform the rest. Therefore, in this work we used the $K_r = 0.6$ and $K_r = 0.7$ EViT-S configurations.

1.3. Selection Of MIL Backbones And Baselines

To facilitate the experiments conducted in this paper, we carefully selected a representative model from each of the CNN and Transformer baselines. This selection process involved an extensive evaluation of each model on the validation set of the ISIC 2019 dataset, focusing on the binary classification problem of melanoma (MEL) versus nevus (NV). The results of this evaluation are summarized in table 1.

	Baseline models	**ISIC 2019**		
		BA	R-MEL	R-NV
CNN	RN-18	88.6	83.8	93.4
	RN-50	88.9	82.6	95.1
	VGG-16	87.7	83.6	91.8
	DN-169	89.1	83.2	95.0
	EN-B3	**90.7**	**85.5**	**95.8**
ViT	ViT-S	91.3	86.8	95.8
	ViT-B	90.6	85.3	95.8
	DEiT-S	**91.7**	86.7	**96.7**
	DEiT-B	91.7	**87.2**	96.2

Table 1. Evaluation results of a set of baseline models on the validation set of the ISIC 2019 dataset. The baseline models include different architectures, including RN-18, RN-50, VGG-16, DN-169, EN-B3 from the CNN-based category, and ViT-S, ViT-B, DEiT-S, DEiT-B from the Transformer-based category. The evaluation is performed for the binary classification problem of melanoma versus nevus.

1.4. Multi-class MIL Complementary Material

A detailed summary of the results for the multi-class MIL is shown in Table 2.

1.5. Additional MIL Heatmaps

Figure 3 provides a visual representation of the different visualizations produced by the *instance-level* MIL model using the **top-k average pooling** operator. In this case, the last row shows the gradients associated with the patches classified as nevus.

Table 2. Results of multi-class image classification on the ISIC 2019 validation set. '**I-1**' and '**I-2**' denote the first and second *instance-level* approaches, respectively, and '**E**' denotes the *embedding-level* approach.

	Models		**ISIC 2019**								
			BA	R-AK	R-BCC	R-BKL	R-DF	R-MEL	R-NV	R-SCC	R-VASC
	EN-B3		82.2	71.1	87.8	79.4	81.3	76.5	91.6	72.2	98.0
	DEiT-S		83.6	72.3	90.5	**82.7**	87.5	**80.3**	92.1	65.1	98.0
EViT	Kr = 0.6		83.6	71.7	89.0	80.4	**93.8**	77.8	91.3	66.7	98.0
EViT	Kr = 0.6		**84.3**	78.6	**90.7**	80.4	87.5	75.6	**93.5**	69.8	98.0
MIL-EN-B3	**I-1**	Max	74.1	57.2	81.5	74.3	77.1	74.6	77.8	61.9	88.2
MIL-EN-B3	**I-1**	Topk	78.4	72.8	78.8	75.6	79.2	74.9	87.3	68.3	90.2
MIL-EN-B3	**I-1**	Avg	79.9	71.7	86.4	78.1	83.3	76.8	84.1	62.7	96.1
MIL-EN-B3	**I-2**	Max	76.4	72.3	82.4	75.4	72.9	66.7	80.0	65.1	96.1
MIL-EN-B3	**I-2**	Topk	76.2	75.7	77.1	65.5	79.2	71.0	82.5	66.7	92.2
MIL-EN-B3	**I-2**	Avg	77.5	68.2	86.5	74.1	77.1	73.5	81.2	69.0	90.2
MIL-EN-B3	**E**	Max	72.3	55.5	77.6	73.7	68.8	66.9	79.0	70.6	86.3
MIL-EN-B3	**E**	Topk	78.9	66.5	86.3	79.6	81.3	74.2	84.1	66.7	92.2
MIL-EN-B3	**E**	Avg	77.6	68.8	84.2	77.3	77.1	77.4	80.7	63.5	92.2
MIL-DEiT-S	**I-1**	Max	82.2	72.8	88.7	81.1	89.6	80.1	82.7	64.3	98.0
MIL-DEiT-S	**I-1**	Topk	81.7	78.0	84.0	73.0	66.7	78.4	88.4	68.3	90.2
MIL-DEiT-S	**I-1**	Avg	81.6	76.3	85.7	72.0	89.6	75.4	88.6	69.0	96.1
MIL-DEiT-S	**I-2**	Max	75.4	70.5	80.9	75.8	72.9	66.6	83.1	61.1	92.2
MIL-DEiT-S	**I-2**	Topk	79.0	74.0	84.5	75.6	72.9	76.3	87.2	65.1	96.1
MIL-DEiT-S	**I-2**	Avg	82.6	79.2	87.8	76.2	87.5	77.9	91.1	66.7	94.1
MIL-DEiT-S	**E**	Max	82.4	**80.9**	87.8	75.1	83.3	76.4	89.0	**74.6**	92.2
MIL-DEiT-S	**E**	Topk	82.2	73.4	83.6	74.3	93.8	75.6	92.2	69.1	96.1
MIL-DEiT-S	**E**	Avg	82.6	70.5	88.7	79.8	89.6	75.1	91.5	65.9	**99.9**

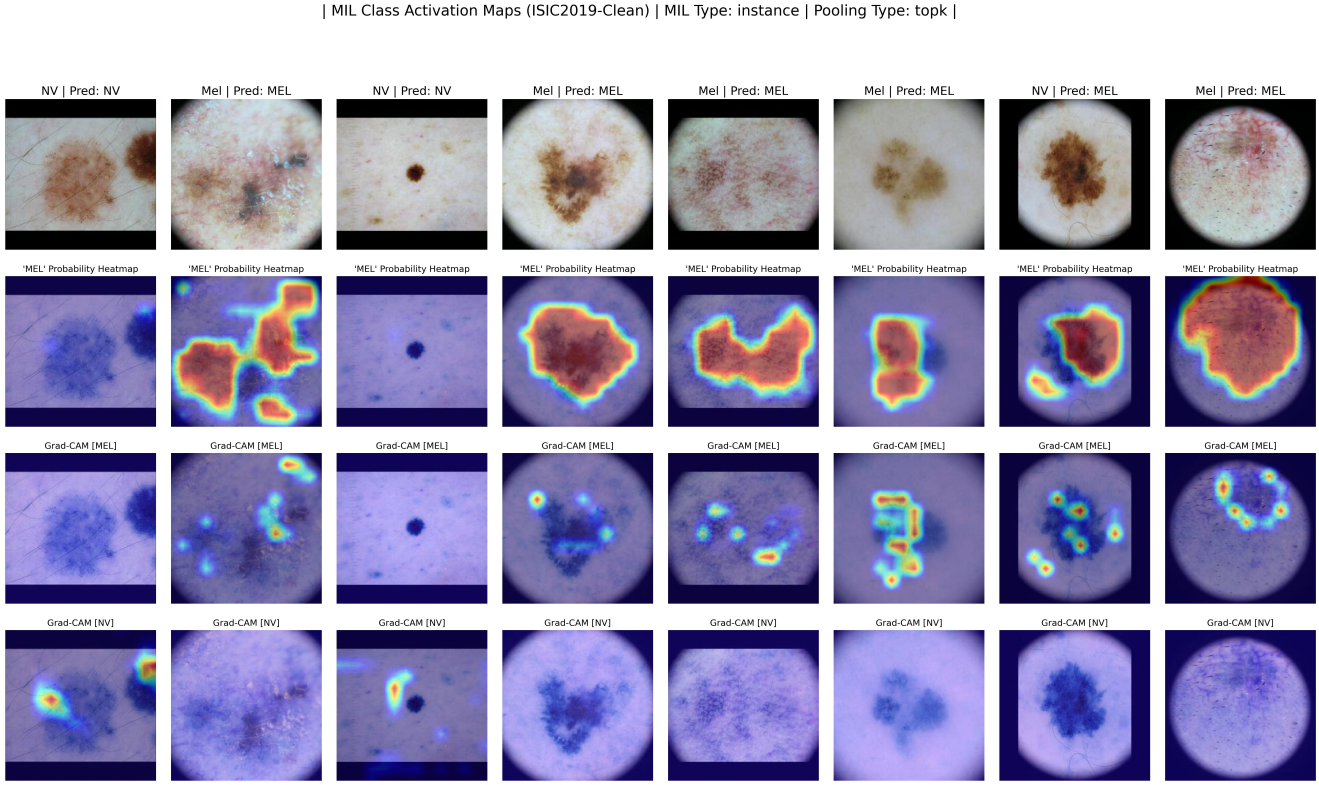

Figure 3. Visualization of the heatmaps generated by the MIL classifier, specifically the *instance-level* MIL model using the **top-k average pooling** operator. The backbone used for the MIL model is the RN-18. The images are taken from the validation set of the ISIC 2019 dataset, and belong to the binary problem of melanoma vs. nevus. The Figure shows the input images in the first row, followed by the patch probability heatmap for the melanoma class in the second row. The third row shows the gradient heatmap for each patch. In this case, the last row shows the gradients with respect to the patches that the model predicted to be nevus.