Supplementary Material: Feature Selection Gates with Gradient Routing for Endoscopic Image Computing

Giorgio Roffo [1]✉, Carlo Biffi [1]
Pietro Salvagnini [1], and Andrea Cherubini [1,2]✉

[1] Cosmo Intelligent Medical Devices, Dublin, Ireland
[2] Milan Center for Neuroscience, University of Milano–Bicocca, Milano, Italy
{*groffo, acherubini*}*@cosmoimd.com*

S1. FSG Weight Distributions in the ViT Model:

Our research highlights the adaptability of Vision Transformer (ViT) architectures enhanced with Feature Selection Gates (FSG) for tasks like CIFAR-100 classification and polyp sizing. CIFAR-100 is a classification task on natural images, while polyp sizing is a regression task on medical images. FSG, which are online feature re-weighting gates, act as a sort of hard-attention mechanism by scaling irrelevant dimensions and embeddings. This attenuates and reduces the model parameters without removing them. FSG dynamically adjusts feature importance based on task requirements, showcasing effective feature selection.

In CIFAR-100 classification, FSG prioritizes a range of features from basic (edges, lines) in early layers to complex (patterns, objects) in deeper layers. Weight distributions shift towards higher significance (initially $\mu = 0.84$, $\sigma = 0.13$, final layers $\mu = 0.92$, $\sigma = 0.07$), reflecting the need for comprehensive feature integration to classify diverse objects accurately. For polyp sizing, a regression task, FSG weights remain uniformly distributed ($\mu \approx 0.5$, σ from 0.026 to 0.029), similar to the model's initial state. This uniformity suggests equitable feature consideration, essential for size and shape differentiation, akin to regularization methods like L1 and L2 that prevent overfitting.

The different FSG weight behaviors in these tasks are due to the nature of the tasks themselves. In classification tasks like CIFAR-100, the need to distinguish between many classes requires FSG to emphasize a wide range of features, aligning with hierarchical feature learning where early layers capture generic features, and deeper layers capture task-specific details. In regression tasks like polyp sizing, the goal is to predict a continuous value based on subtle differences in features such as size and shape. The uniform FSG weights ensure a balanced consideration of all features, minimizing regression error without overemphasis on specific features, similar to the effect of regularization techniques that prevent overfitting. In summary, FSG's adaptive feature moderation optimizes performance by emphasizing critical features in classification tasks and maintaining balanced feature integration in regression tasks, aligning with task-specific requirements. Additionally, optimizing FSG parameters separately from the main model allows tailored learning rates and gradient clippings, thus enhancing FSG networks' training efficiency.

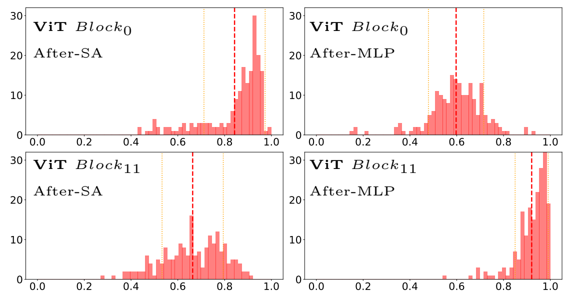

Fig. 1. CIFAR-100: FSG-GR weight distributions in the ViT.

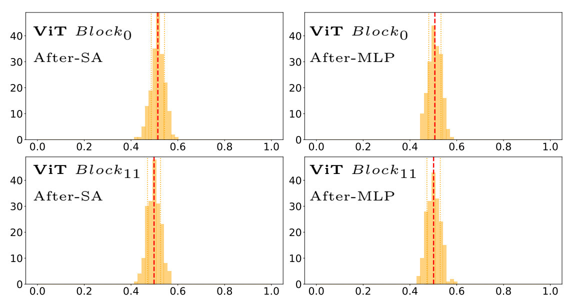

Fig. 2. Polyp Sizing: FSG-GR weight distributions in the ViT.

Table 1. Experimental Settings for Models with and without FSG

Category	**Parameter**	**Experimental Settings**	
		With FSG	**Without FSG**
Preprocessing	Image Resizing	384×384 pixels	
	Normalization Means	0.32239652, 0.22631808, 0.17500061	
	Normalization STDs	0.31781745, 0.2405859, 0.19327126	
Augmentation	Techniques	Random Rotations, Color Jittering, Gaussian Noise	
Target Norm.	Range	$[-1, +1]$	
	Min-Max Values	0.5 mm, 20.0 mm	
Data Balancing	Frame Selection	40-128 Frames per Video	
Loss Function	Type	Weighted Huber	Weighted Huber
	Parameters Model	Weight [3,5], Threshold [5,10]	Weight [3,5], Threshold [5,10]
	Parameters FSG	Weight [5,10], Threshold [5,10]	N.A.
Training	Optimizer	Adam	
	Learning Rate	1×10^{-2}-1×10^{-5}	1×10^{-3}
	Weight Decay	1×10^{-2}-1×10^{-8}	1×10^{-5}
	Gradient Clipping	64.0-128.0	5.0-8.0