Token Pruning using a Lightweight Background Aware Vision Transformer

Sudhakar Sah, Ravish Kumar, Honnesh Rohmetra, Ehsan Saboori
Deeplite, Toronto, Canada

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

High Computational Complexity of ViTs
- High runtime memory usage
- High latency

Token Pruning
- Technique to remove less important tokens based on their relevance
- Memory reduction, latency reduction

Contribution:
- A novel Background Aware Vision Transformer (BAViT) approach, capable of separating foreground (FG) and background tokens (BG) efficiently.
- A modified Accumulative Cross Entropy Loss function for classification.
- An improved throughput of object-detection models by 30-40% with minimal accuracy drop that can be regained after finetuning for few epochs.

Motivation

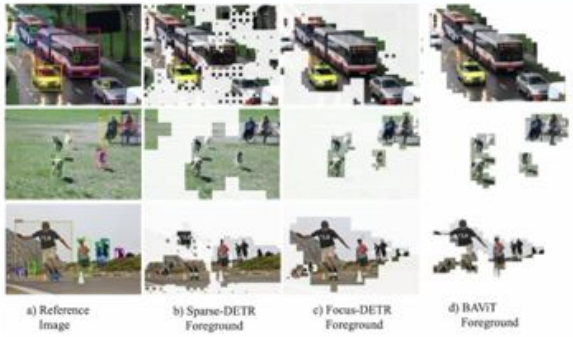

Figure 1: Comparison of BAViT with others

- The number of input tokens in ViTs has the quadratic complexity which limits their application specially in resource-constrained hardware.
- Methods like Sparse DETR and Focus DETR [1] have proven that token pruning can improve latency and throughput of object detection models.
- Focusing on efficient pruning without the computational overhead of heavy CNN backbones, making it ideal for edge-device applications.

Figure 2: (Left) Original image, (Center) foreground object area, (Right) 16 ×16 grid with red grids being foreground

BAViT Methodology

Annotation Formation Each token is labeled as FG if the token has more than 50% overlapping area with any foreground bounding box else it is labeled as BG, forming an M-dimensional annotation vector for input images, primarily used for training the BAViT model.

$$\mathbf{L_i} = \begin{cases} 1 & \{\text{if } J(P_i, B_j) \geq \tau \\ 0 & \{\text{if } J(P_i, B_j) < \tau \end{cases} \quad (1) \qquad \mathbf{J(P_i, B_j)} = \frac{|P_i \cap B_j|}{|P_i \cup B_j|} \quad (2)$$

Accumulative Cross Entropy Loss Instead of using loss from a single classification token, the proposed method calculates an Accumulative Cross Entropy Loss (I_{acc}) by summing individual token losses across all tokens, improving classification accuracy for each image patch.

$$\mathbf{L_{acc}} = -\frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{c=1}^{C} y_{i,j,c} \log(\hat{y}_{i,j,c}) \quad (3)$$

Background Aware ViT (BAViT) Architecture

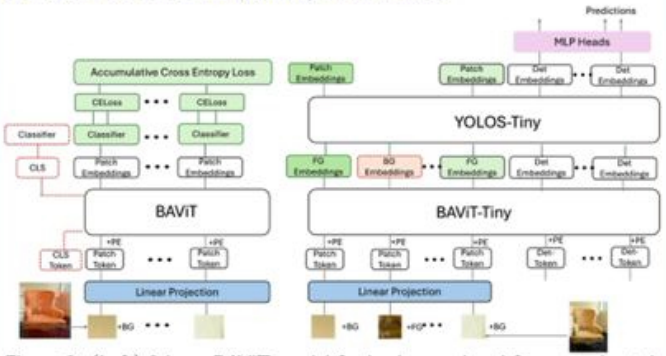

Figure 3: (Left) 2 layer BAViT model for background and foreground patch classification. (Right) BAViT as a pre-processing step to YOLOS to perform end-to-end object detection.

Model Training
- BAViT model trained forFG/BG classification using Accumulative Cross Entropy Loss (small and large models trained using VOC/COCO datasets)
- BAViT attached to YOLOS to classify tokens as FG/BG and processing only FG tokens to reduce computation efficiently.

TABLE 1 : Models Accuracy Table

Model	Depth	Dataset	Accuracy(%)
BAViT-small	2	Pascal-VOC	75.93
BAViT-large	10	Pascal-VOC	88.79
BAViT-small	2	MS-COCO	70.88
BAViT-large	10	MS-COCO	80.57

Experimental Results

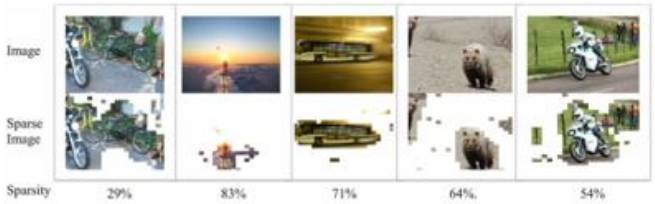

Figure 4 : FG/BG token classification (16x16) on COCO images

- The BAViT model significantly reduces tokens (up to 24% with 34% sparsity) in ViT-based object detection tasks like YOLOS, improving efficiency for edge use cases with minimal mAP drop, which can be recovered.
- BAViT provides a lightweight, plug-and-play solution, offering configurable sparsity for optimal accuracy-latency trade-offs for edge devices.

TABLE 2: Token reduction using BAViT on YOLOS-tiny model. Total Tokens for a 384 resolution image (BAViT-small: 1152 & YOLOS: 12288)

			Number of Tokens		
Model	**Sparsity %age**	**mAP**	**YOLOS Pruned**	**YOLOS +BAViT**	**Percent Reduction**
BAViT+YOLOS	46%	20.00	6635	7787	36.63%
BAViT+YOLOS	43%	21.50	7004	8156	33.63%
BAViT+YOLOS	40%	22.50	7372	8524	30.63%
BAViT+YOLOS	39%	22.70	7495	8647	29.63%
BAViT+YOLOS	37%	23.80	7741	8893	27.63%
BAViT+YOLOS	35%	**24.40**	7987	9139	25.63%
BAViT+YOLOS-F	35%	**26.60**	7987	9139	25.63%
BAViT+YOLOS	32%	25.00	8355	9507	22.60%
BAViT+YOLOS	29%	25.90	8724	9876	19.60%
BAViT+YOLOS	5%	27.70	11673	12825	-4.37%
BAViT+YOLOS	2%	28.60	12042	13194	-7.38%
BAViT+YOLOS	0%	28.80	12288	13440	-9.40%

Summary

- BAViT-small model prunes 25% of YOLOS-tiny tokens, with a 3% mAP drop recoverable to under 2% via sparse token fine-tuning for 30 epochs.
- It offers a low-cost, edge-friendly alternative to methods like Focus DETR, supporting joint training and adaptive sparsity based on image complexity.

References:

[1] Hailin Hu Dehua Zheng, Wenhui Dong. Less is more: Focus attention for efficient detr.
[2] A Kolesnikov A Dosovitskiy, L Beyer. An image is worth 16x16 words: Transformers for image recognition at scale.