Leveraging *Sparsity* for Efficient Inference of *High-Resolution* Vision Foundation Models

Xin Xu, Jason Kuen, Brian L. Price, Kangning Liu, Zijun Wei, Yu-Xiong Wang

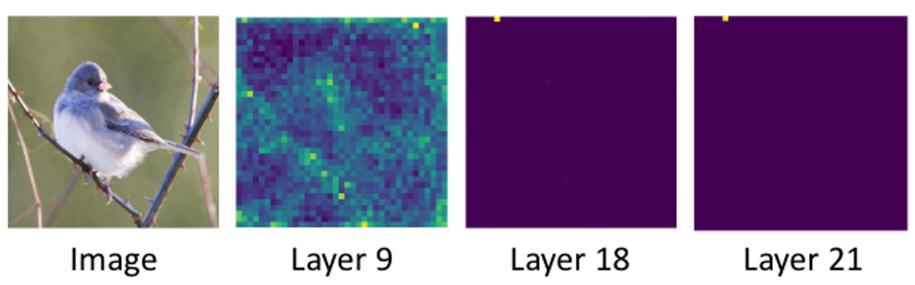
Resolution Scaling

Resolution Scaling: *Higher resolution, better performance*.

Key Question: How to address attention's computational bottlenecks?

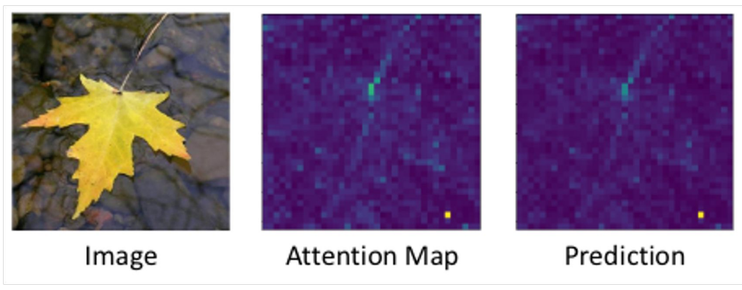
Observations on Sparsity

Visualizations of Attention Map in DINOv2 layers:



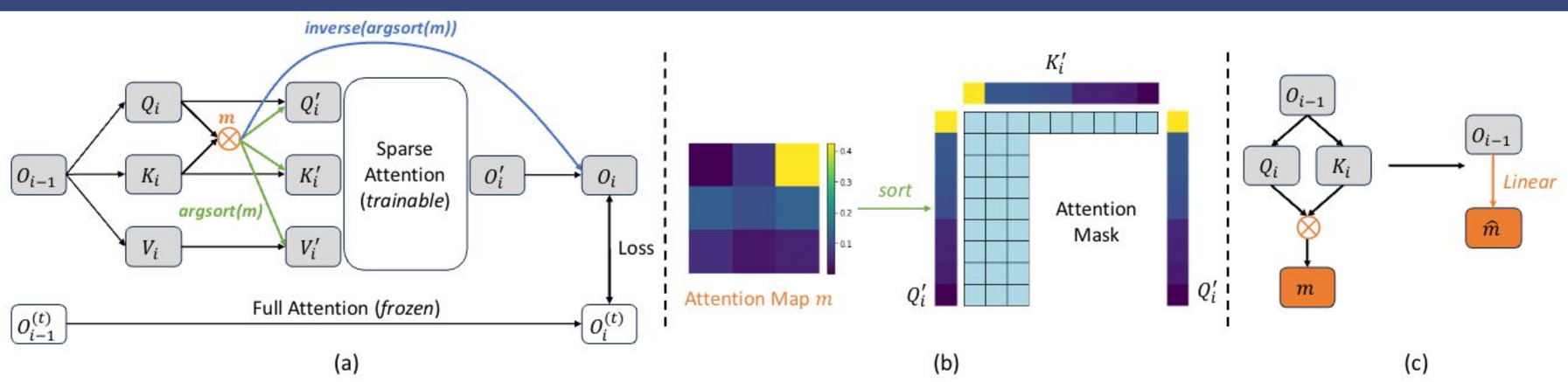
Observation #1: Activation concentrates on a few tokens *in deeper layers*.

Observation #2: Sparsity patterns *persist* across layers (e.g. layer 18 and layer 21).



Observation #3: A *lightweight* projector can accurately predict sparsity patterns, eliminating costly matrix multiplications.

SVE: Unlocking Sparse Vision Encoder



(a) We use *layer-wise* distillation to train a student model that adapts with sparse attention guided by the teacher's "ground-truth" sparsity pattern.

(b) To compute efficiently, we sort tokens based on *activations*, compute sparse attention, and restore the original order.

(c) Replace heavy Matmul with a lightweight linear projector for sparsity prediction.

Experiments

Models		ImageNet-1K [13]	ADE20K [48, 49]	Pascal VOC [30]	TFLOP	Speedup
DINOv2 [32] (1K)	Pre-trained	84.47	45.54	75.37	6.2	1.76
	SVE (Ours)	84.17	44.70	75.15	**5.5** (-11%)	**2.03** (+15%)
CLIP [35] (1K)	Pre-trained	73.99	-	-	5.8	1.97
	SVE (Ours)	72.76	-	-	**4.6** (-21%)	**2.58** (+31%)
SAM2 [36] (1K)	Pre-trained	-	61.22	74.22	0.6	14.62
	SVE (Ours)	-	60.81	74.50	**0.5** (-17%)	**16.80** (+15%)

SVE consistently improves efficiency without sacrificing performance, which proves to be a reliable post-training optimization technique.

Models	Resolution	HQSeg-44K [34]	Speedup
DINOv2 [32]	518x518*	55.95	1.1x
	1036x1036	**58.24**	**1.2x**
SAM2 [36]	1024x1024*	86.96	1.2x
	2048x2048	**87.63**	**1.8x**

The power of resolution scaling, and the effect of SVE with resolution scaling.