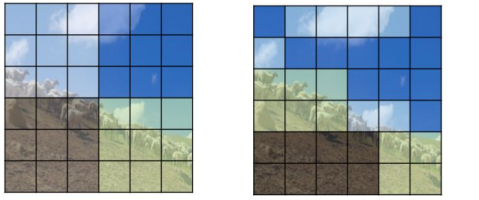
BOAT: Bilateral Local Attention Vision Transformer

Tan Yu, Gangming Zhao, Ping Li, Yizhou Yu

香 港 大 學
THE UNIVERSITY OF HONG KONG



(a) Image-space Local Attention (b) Feature-space Local Attention

Figure 1: The image-space local attention versus the feature-space local attention.

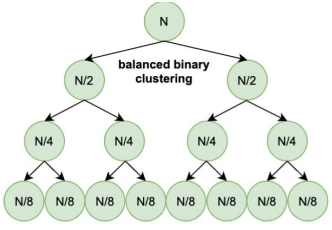


Figure 4: Example of balanced hierarchical clustering. In this example, the number of hierarchical levels is 3. There are $2^3 = 8$ clusters in the bottom level.

Algorithm 1: Balanced Binary Clustering.

Input: Tokens $\{\mathbf{t}_i\}_{i=1}^{2m}$ and the iteration number, T.
Output: Two clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$.

Initialize centroids $\mathbf{c}_1 = \frac{\sum_{i=1}^{m} \mathbf{t}_i}{m}$, $\mathbf{c}_2 = \frac{\sum_{i=m+1}^{2m} \mathbf{t}_i}{m}$
while *n_iter* $\in [1, T]$ **do**
  for $i \in [1, 2m]$ **do**
    $r_i = \frac{s(\mathbf{t}_i, \mathbf{c}_1)}{s(\mathbf{t}_i, \mathbf{c}_2)}$
  $[i_1, \cdots, i_{2m}] = \text{argsort}([r_1, \cdots, r_{2m}])$
  $\mathcal{C}_1 = \{\mathbf{t}_{i_j}\}_{j=1}^{m}$, $\mathcal{C}_2 = \{\mathbf{t}_{i_j}\}_{j=m+1}^{2m}$
  $\mathbf{c}_1 = \frac{\sum_{\mathbf{c} \in \mathcal{C}_1} \mathbf{c}}{m}$, $\mathbf{c}_2 = \frac{\sum_{\mathbf{c} \in \mathcal{C}_2} \mathbf{c}}{m}$

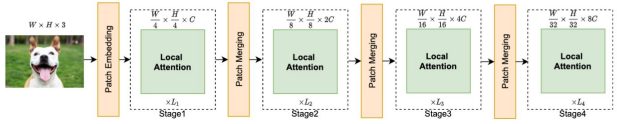


Figure 2: Architecture of Bilateral lOcal Attention Vision Transformer (BOAT).

Figure 3: Architecture of Bilateral Local Attention (BLA) Block.

$$\mathcal{T}_{\text{ISLA}} = \mathcal{T}_{\text{in}} + \text{ISLA}(\text{LN}(\mathcal{T}_{\text{in}})).$$

$$\mathcal{T}_{\text{FSLA}} = \mathcal{T}_{\text{ISLA}} + \text{FSLA}(\text{LN}(\mathcal{T}_{\text{ISLA}})).$$

$$\mathcal{T}_{\text{out}} = \mathcal{T}_{\text{FSLA}} + \text{MLP}(\text{LN}(\mathcal{T}_{\text{FSLA}})).$$

	BOAT-Swin-T	BOAT-Swin-S	BOAT-Swin-B
stage 1	concat 4 × 4 [win. sz. 7×7, dim 96, head 3 Swin-Block win. sz. 7×7, dim 96, head 3 BLA-Block] × 1	concat 4 × 4 [win. sz. 7×7, dim 96, head 3 Swin-Block win. sz. 7×7, dim 96, head 3 BLA-Block] × 1	concat 4 × 4 [win. sz. 7×7, dim 128, head 4 Swin-Block win. sz. 7×7, dim 128, head 4 BLA-Block] × 1
stage 2	concat 2 × 2 [win. sz. 7×7, dim 192, head 6 Swin-Block win. sz. 7×7, dim 192, head 6 BLA-Block] × 1	concat 2 × 2 [win. sz. 7×7, dim 192, head 6 Swin-Block win. sz. 7×7, dim 192, head 6 BLA-Block] × 1	concat 2 × 2 [win. sz. 7×7, dim 256, head 8 Swin-Block win. sz. 7×7, dim 256, head 8 BLA-Block] × 1
stage 3	concat 2 × 2 [win. sz. 7×7, dim 384, head 12 Swin-Block win. sz. 7×7, dim 384, head 12 BLA-Block] × 3	concat 2 × 2 [win. sz. 7×7, dim 384, head 12 Swin-Block win. sz. 7×7, dim 384, head 12 BLA-Block] × 9	concat 2 × 2 [win. sz. 7×7, dim 512, head 16 Swin-Block win. sz. 7×7, dim 512, head 16 BLA-Block] × 9
stage 4	concat 2 × 2 [win. sz. 7×7, dim 768, head 24 Swin-Block] × 2	concat 2 × 2 [win. sz. 7×7, dim 768, head 24 Swin-Block] × 2	concat 2 × 2 [win. sz. 7×7, dim 1024, head 32 Swin-Block] × 2

Method	size	#para.	FLOPs	Top-1	Method	size	#para.	FLOPs	Top-1
ReGNetY-4G [21]	224	21M	4.0G	80.0					
PVTv2-B2 [27]	224	25M	4.0G	82.0	Focal-T [31]	224	29M	4.9G	82.2
Swin-T [18]	224	29M	4.5G	81.3	BOAT-Swin-T (ours)	224	31M	5.2G	82.3
CSWin-T [8]	224	23M	4.3G	82.7	BOAT-CSWin-T (ours)	224	27M	5.1G	**83.7**
ReGNetY-8G [21]	224	39M	8.0	81.7	PVTv2-B4 [27]	224	62M	10.1G	83.6
Twins-B	224	56M	8.3G	83.2	Shuffle-S [14]	224	50M	8.9G	83.5
NesT-S [37],	224	38M	10.4G	83.3	Focal-S [31]	224	51M	9.1G	83.5
Swin-S [18]	224	50M	8.7G	83.0	BOAT-Swin-S (ours)	224	56M	10.1G	83.6
CSWin-S [8]	224	35M	6.9G	83.6	BOAT-CSWin-S (ours)	224	41M	8.0G	**84.1**
ReGNetY-16G [21]	224	84M	16.0G	82.9	ViT-B/16T [9]	384	86M	55.4G	77.9
DeiT-B [25]	224	86M	17.5G	81.8	T2T-24 [36]	224	64M	14.1G	82.3
TNT-B [10]	224	66M	14.1G	82.8	PiT-B [13]	224	74M	12.5G	82.0
PVTv2-B5 [27]	224	82M	11.8G	83.8	Twins-L	224	99M	14.8G	83.7
Shuffle-B [14]	224	88M	15.4G	84.0	NesT-B [37],	224	68M	17.9G	83.8
Focal-B [31]	224	90M	16.0G	83.8	CrossFormer-L [29]	224	92M	16.1G	84.0
Swin-B [18]	224	88M	15.4G	83.5	BOAT-Swin-B (ours)	224	98M	17.8G	83.8
CSWin-B [8]	224	78M	15.0G	84.2	BOAT-CSWin-B (ours)	224	90M	17.5G	**84.7**

Table 1: Comparison of image classification performance on the ImageNet-1K dataset.

Method	#para.(M)	FLOPs(G)	mIoU(%)	Method	#para.(M)	FLOPs(G)	mIoU(%)
TwinsP-S [6]	55	919	46.2	Twins-S [6]	54	901	46.2
Shuffle-T [14]	60	949	46.6	Focal-T [31]	62	998	45.8
Swin-T [18]	60	945	44.5	BOAT-Swin-T (ours)	62	986	46.0
CSWin-T [8]	60	959	49.3	BOAT-CSWin-T (ours)	64	1012	**50.5**
TwinsP-B [6]	74	977	47.1	Twins-B [6]	89	1020	47.7
Shuffle-S [14]	81	1044	48.4	Focal-S [31]	85	1130	48.0
Swin-S [18]	81	1038	47.6	BOAT-Swin-S (ours)	87	1113	48.4
CSWin-S [8]	65	1027	50.0	BOAT-CSWin-S (ours)	70	1101	**50.6**
TwinsP-L [6]	92	1041	48.6	Twins-L [6]	133	1164	48.8
Shuffle-B [14]	121	1196	49.0	Focal-B [31]	126	1354	49.0
Swin-B [18]	121	1188	48.1	BOAT-Swin-B (ours)	131	1299	48.7
CSWin-B [8]	109	1222	50.8	BOAT-CSWin-B (ours)	121	1349	**50.9**

Table 5: Performance of semantic segmentation on ADE20K. FLOPs are obtained at 512 × 2048 resolution. mIoU is for the single-scale setting. Testing image size is 512 × 512.

Method	#para.(M)	FLOPs(G)	mAP^{box}	mAP^{Mask}
Swin-T	48	267	46.0	41.6
BOAT-Swin-T (ours)	50	306	**47.5**	**42.8**
Swin-S	69	359	48.5	43.3
BOAT-Swin-S (ours)	75	431	**49.0**	**43.8**

Table 6: Performance of object detection on the MS-COCO dataset. FLOPs are obtained at 800 × 1280 resolution.

Method	Reformer	K-means	Ours
Top-1 Accuracy	81.7	81.8	82.3

Table 2: Comparison of image classification accuracy with Reformer and K-means.

Model	BOAT-Swin-T (with FSLA)	Baseline (with ISLA)
Accuracy	82.3	81.5

Table 3: Ablation study on FSLA by replacing FSLA with ISLA.

Overlap	BOAT-CSWin-T	BOAT-CSWin-S	BOAT-CSWin-B
No	83.3%	84.0%	84.5%
Yes	83.7%	84.1%	84.7%

Table 4: Comparison of image classification accuracy between overlapping balanced hierarchical clustering and the non-overlapping version.