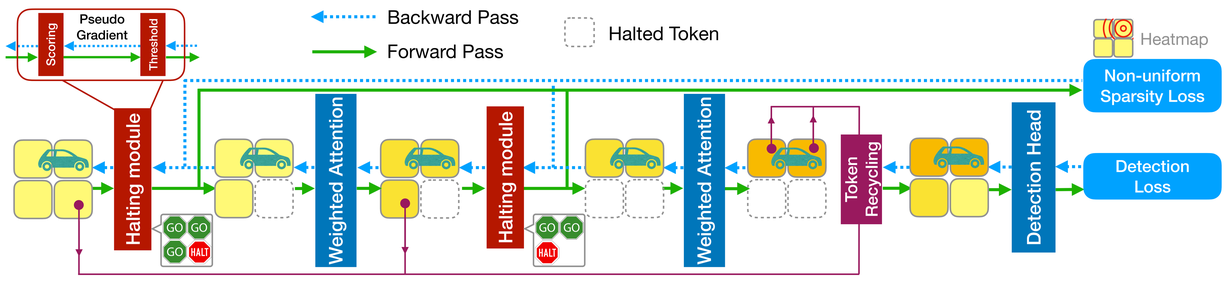
Efficient Transformer-based 3D Object Detection with Dynamic Token Halting

Mao Ye[1,2], Gregory P. Meyer[2], Yuning Chai[2], and Qiang Liu[1]
[1]The University of Texas at Austin
[2]Cruise LLC

Introduction

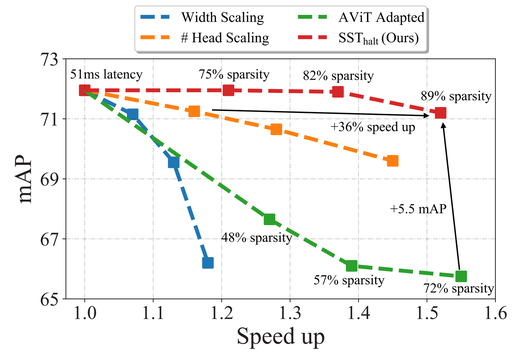
- Transformers have achieved state-of-the-art results in several computer vision tasks.
- However, transformers suffer from high-latency due the quadratic complexity of attention.
- In this work, we propose a method for accelerating transformer-based 3D object detectors by dynamically halting tokens based on their contribution to the detection task.
- As a result, our method significantly improves the Pareto frontier of latency versus accuracy.

Method



- Our method is built upon SST, which is a transformer-based LiDAR object detector.
- First, the input LiDAR point cloud is voxelized, and each voxel is treated as a token.
- Before each layer, a halting module scores the tokens and halts those with a low score.
- During training, we define a pseudo-gradient to back-propagate through the halting module.
- To improve learning of the halting module, we leverage a weighted attention mechanism, which weights the attention given to each token based on their score.
- After the last attention layer, halted tokens are recycled and forwarded to the detection head.
- The whole network is trained end-to-end using a detection loss and a sparsity loss.

Efficiency vs. Accuracy



When compared to other model scaling approaches, our method provides the best efficiency and accuracy trade-off.

Detection Performance

WOD Validation Set

Method	Vehicle APH L2	Pedestrian APH L2	Cyclist APH L2
PointPillar	63.1	50.3	59.9
PV-RCNN	68.4	57.6	64.0
RangeDet	63.6	63.9	62.1
LiDAR R-CNN	67.9	51.7	64.4
CenterPoint	67.5	57.9	—
RSN	65.5	63.7	—
SST	65.1	61.7	—
SWFormer	68.8	64.9	—
SST$^{++}_{halt}$ (Ours)	**69.0**	**66.5**	**66.0**

We leverage the latency savings provided by our method to improve the performance of SST while maintaining its runtime.

Visualizations