Quadtree Generating Networks: Efficient Heirarchical Scene Parsing with Sparse Convolutions

Kashyap Chitta[1], Jose M. Alvarez[2], Matrial Hebert[3]

[1] MPI-IS and University of Tübingen | [2] NVIDIA | [3] Carnegie Mellon University

Semantic Segmentation

Problem

- Segmentation is a memory-intensive task due to **high resolution**, **dilation-based operations** that maintain high resolution activations, and **quadratic scaling**
- Leads to (1) poor training due to low batch sizes, (2) latency at inference

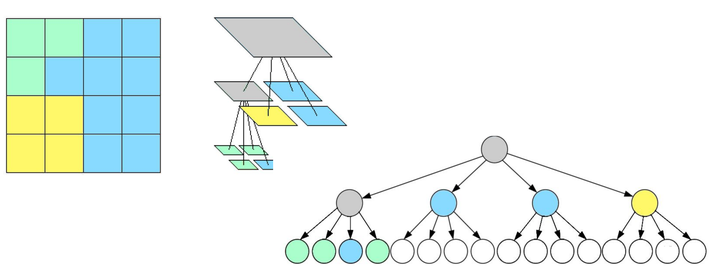
- Goal: improve performance-memory trade-offs with **sparse output representations**

Quadtrees

- Hierarchical representation of 2D grids
- Advantage: Memory scales sub-quadratically, based on the number of pixels at class boundaries

Segmentation Maps as Quadtrees

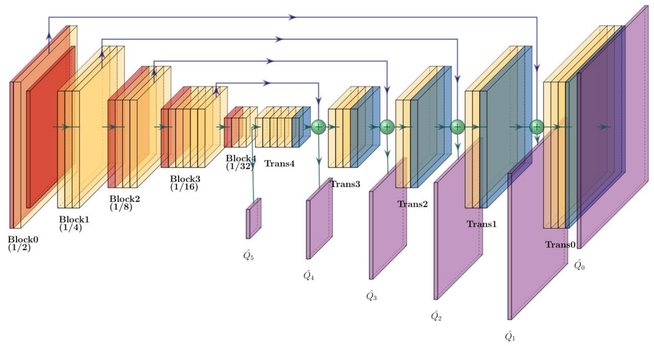
- Starting from full-resolution, recursively group 4 neighboring pixels (children) to a single node (parent)
- **Composite class** assigned if children belong to different classes (grey in illustration)



Quadtree Generating Networks

Architecture

- Encoder-decoder with skip connections and no dilated convolutions



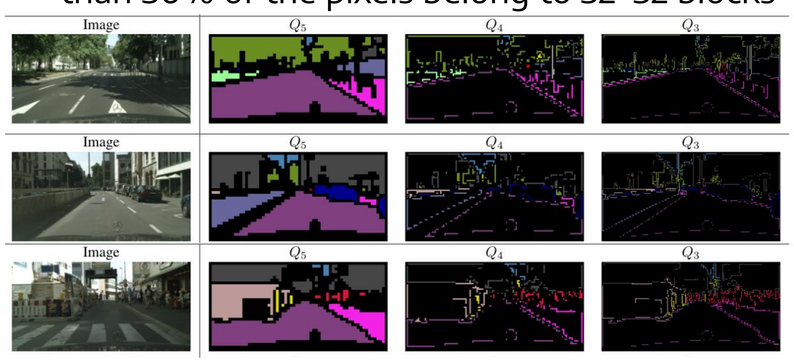
- Decoder activations stored as **hash tables** (x, y, feature)
- **Prediction layers** at each block that control propagation

Propagation Scheme

- If prediction = composite class, activation propagated to next layer, else set to zero
- Can adjust propagation to trade-off performance and memory consumption

Datasets

- In Cityscapes, SUN-RGBD and ADE20k, more than 50% of the pixels belong to 32*32 blocks



Results

Results on Cityscapes

- QGN-All: high-memory propagation scheme with all pixels propagated
- QGN-PC: low-memory with only composite class propagated
- Good trade-off between accuracy and memory

Model	Memory (GB)	Compute (TFLOPS)	mIoU
DRN-C-42[1]	3.77	1.07	70.9
DRN-D-105[1]	15.15	1.91	75.6
DeepLabv3[2]	14.27	1.97	79.3
QGN-All (Ours)	**5.85**	**0.48**	**78.2**
QGN-PC (Ours)	**3.66**	**0.25**	**73.0**

Results on ADE20k

- QGN can be combined with recent backbone architectures as a drop-in replacement for dilated convolutions

Model	Memory (GB)	Accuracy	mIoU
PSPNet[3]	~2.5	81.39	43.29
PSPNet + QGN	**~1.2**	**81.67**	**43.91**

Conclusions

- Segmentation predictions can be made hierarchically using quadtree representations
- Results competitive to state-of-the-art with **2x-4x less memory consumption**
- Flexible approach that can be **adapted at inference** without retraining

References

[1] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. *Dilated Residual Networks.* In CVPR, 2017.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. *Rethinking Atrous Convolution for Semantic Image Segmentation.* ArXiv, 2017.

[3] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. *Pyramid Scene Parsing Network.* In CVPR, 2017.

Max Planck Institute for Intelligent Systems · NVIDIA · Carnegie Mellon University