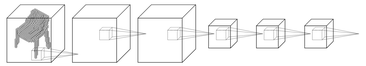
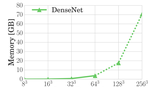
OctNet: Learning Deep 3D Representations at High Resolutions

Gernot Riegler[1], Ali Osman Ulusoy[2,3], Andreas Geiger[2,4]
[1]Graz University of Technology [2]MPI for Intelligent Systems Tübingen [3]Microsoft [4]ETH Zürich

Motivation

In 3D convolutional networks memory requirements increase cubically wrt. input resolution



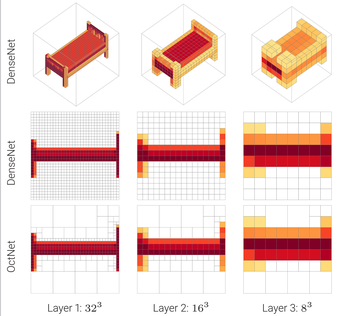


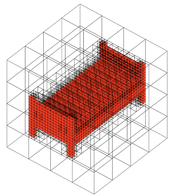
3D data is usually very sparse

Mesh 16^3, 4.19% 64^3, 1.06% 256^3, 0.31%

Idea

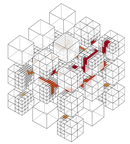
Use space partitioning data-structure within convolutional networks





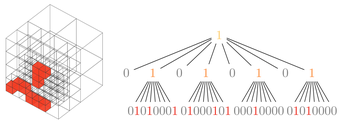
OctNet Representation

Data structure: grid of shallow octrees [1]



- Shallow octree have fixed depth
- Efficiently encoded with bit-string
- Bit indicates if node is split, or not
- Fast neighbour address resolution via bit operations

1 01010101 00000000 01010001 00000000 01000101 00000000 00010000 00000000 01010000



$$\mathrm{pa}(i) = \left\lfloor \frac{i-1}{8} \right\rfloor, \quad \mathrm{ch}(i) = 8 \cdot i + 1,$$

$$\mathrm{data\_idx}(i) = 8 \underbrace{\sum_{j=0}^{\mathrm{pa}(i)-1} \mathrm{bit}(j) + 1}_{\text{\#nodes above i}} - \underbrace{\sum_{j=0}^{i-1} \mathrm{bit}(j)}_{\text{\#split nodes pre i}} + \underbrace{\mathrm{mod}\,(i-1, 8)}_{\text{offset}}$$

OctNet Pooling

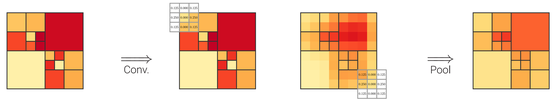
Combines shallow octrees
Data on finest resolution is pooled

$$O^{\mathrm{out}}[i,j,k] = \begin{cases} O^{\mathrm{in}}[2i,2j,2k] \text{ if } \mathrm{vxd}(2i,2j,2k) < 3 \\ P \qquad\qquad \text{else} \end{cases}$$

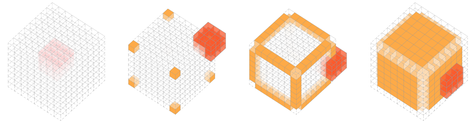
OctNet Convolution

Convolution pools data in larger octree cells

$$O^{\mathrm{out}}[i,j,k] = \underset{(\tilde{i},\tilde{j},\tilde{k}) \in \Omega[i,j,k]}{\mathrm{pool\_voxels}}\,(T_{\tilde{i},\tilde{j},\tilde{k}}) \quad \text{with} \quad T_{i,j,k} = \sum_{l=0}^{L-1} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W_{l,m,n} \cdot O^{\mathrm{in}}[\hat{i},\hat{j},\hat{k}]$$



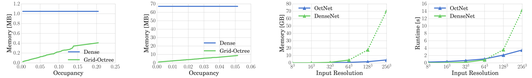
Efficient implementation



Experiments - 3D Classification

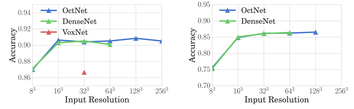
Evaluation on ModelNet dataset [2]

- Memory & Runtime



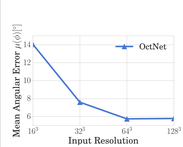
64^3 256^3 Memory Runtime

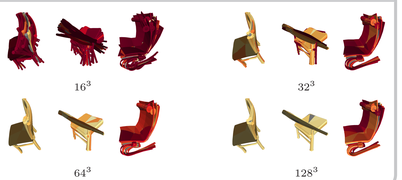
- Accuracy



ModelNet10 ModelNet40

Experiments - 3D Orientation Estimation

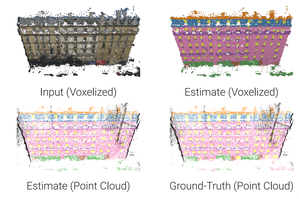




Experiments - 3D Semantic Segmentation

Evaluation on Varcity dataset [3]

	Average	Overall	IoU
Riemenschneider et al.[3]	-	-	42.3
Martinovic et al.[4]	-	-	52.2
Gadde et al.[5]	68.5	78.6	54.4
OctNet 64^3	60.0	73.6	45.6
OctNet 128^3	65.3	76.1	50.4
OctNet 256^3	**73.6**	**81.5**	**59.2**



Input (Voxelized) Estimate (Voxelized)
Estimate (Point Cloud) Ground-Truth (Point Cloud)

Conclusion

OctNet makes high resolution 3D convolutional networks tractable
Can be extended to generate high resolution 3D output [6]
Code is online: `https://github.com/griegler/octnet`

References

[1] Andrew Miller et al. "Real-time rendering and dynamic updating of 3-d volumetric data". In: *GPGPU*. 2011.
[2] Zhirong Wu et al. "3D ShapeNets: A deep representation for volumetric shapes". In: *CVPR*. 2015.
[3] Hayko Riemenschneider et al. "Learning Where to Classify in Multi-view Semantic Segmentation". In: *ECCV*. 2014.
[4] Andelo Martinovic et al. "3D All The Way: Semantic Segmentation of Urban Scenes from Start to End in 3D". In: *CVPR*. 2015.
[5] Raghudeep Gadde et al. "Efficient 2D and 3D Facade Segmentation using Auto-Context". In: *arXiv.org* 1606.06437 (2016).
[6] Gernot Riegler et al. "OctNetFusion: Learning Depth Fusion from Data". In: *arXiv preprint arXiv:1704.01047* (2017).