R-COVNET: RECURRENT NEURAL CONVOLUTION NETWORK FOR 3D OBJECT RECOGNITION

Danielle Tchuinkou Kwadjo, Christophe Bobda
CSCE Department
Smartest Lab

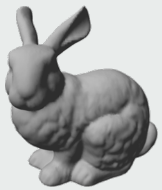
Agenda

I. Introduction
II. Related work
III. Our Approach
IV. Results
V. Conclusion

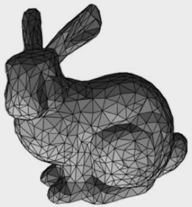
Introduction

- Object recognition with 2D features performs poorly under:
  - Various lighting conditions,
  - Texture,
  - Orientation.

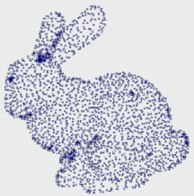
- These problems can be overcome under 3D environments:
  - Descriptors
  - Grid based: mesh and voxel
  - Points based



Volumetric



Mesh

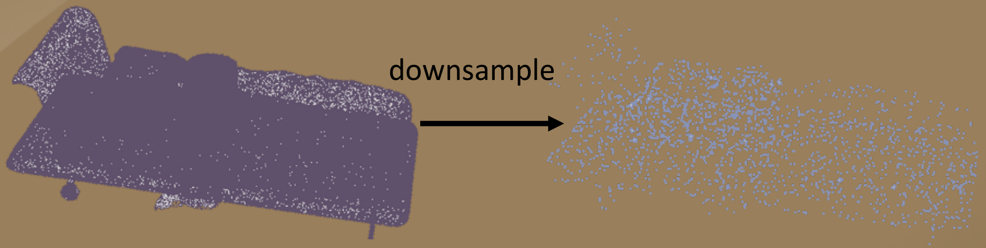


Point Cloud

UNIVERSITY OF ARKANSAS

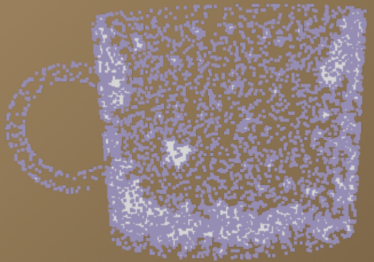
Introduction

- Basic Architecture in literature:
    - CNN with fixed input size (2048): downsample input



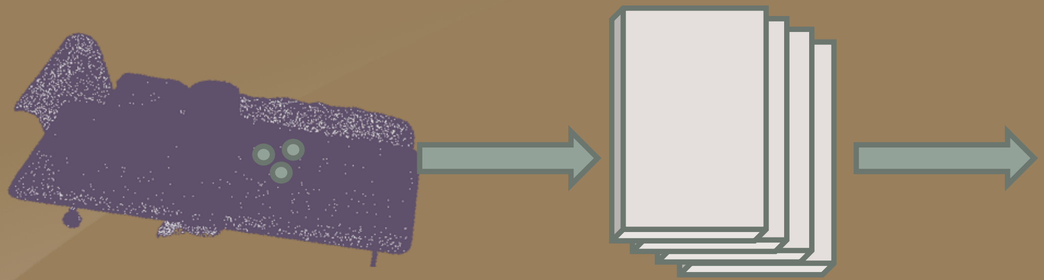
bed of **17,244** points

bed of **2048** points



Cup of **2048** points

Introduction



bed of 17,244 points

R-CovNet

Cup: 0.02

Chair: 0.04

Bed: 0.92 ✓

Table: 0.3

Television: 0.01

....

We need

Network capable of handling input of different size

+

Network able to learn high level features

Question!

- Can we effectively learn features from point clouds without any **preprocessing** and **sampling**?

UNIVERSITY OF ARKANSAS

Question!

- Can we effectively learn features from point clouds without any **preprocessing** and **sampling**?

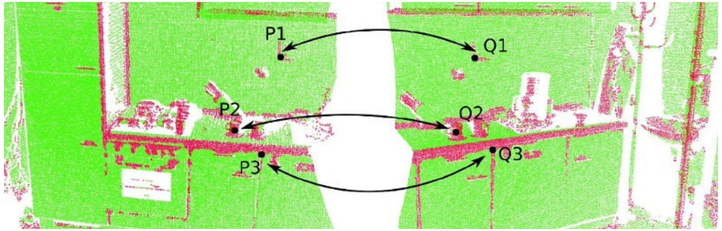
Idea

Deep learning architecture with variant input size, invariant to permutation, robust to long sequence of data and able to learn high level features.

Previous Works: Descriptors

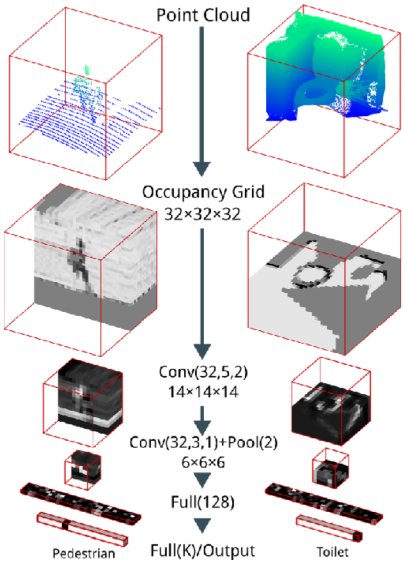
Descriptors

- Build a dataset of descriptors (PFH, SHOT, …) from point clouds
- From an input, find a set of correspondence with the dataset.
- Drawback: extraction of descriptors of the matching algorithms are too computational expensive.



Previous Works: Grid based networks

VoxNet

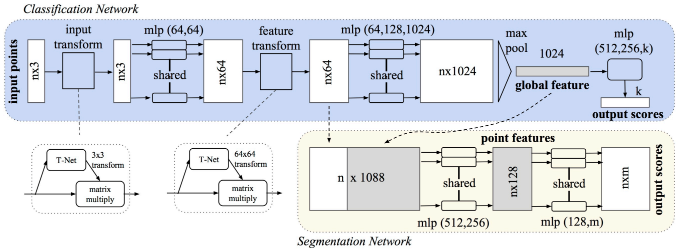


Voxnet [Maturana et al. IROS 2015]

- Fully volumetric approach
- Preprocessing: down-sample input voxel to a fixed size (32x32x32)
- Integrates a volumetric occupancy grid representation with a supervised 3D Convolutional Neural Network
- Suffers from performance loss. Operations on mesh or Voxel are computational expensive.

Previous Works: Points based networks

PointNet

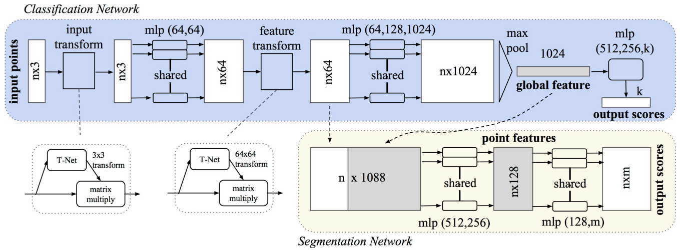


PointNet [Qi et al. CVPR 2017]

- Feed the network directly with points and without prior transformation.
- Takes n points (x, y, z) as input.
- Applies input and feature transformations.
- The output is classification scores for k classes.

Previous Works: Points based networks

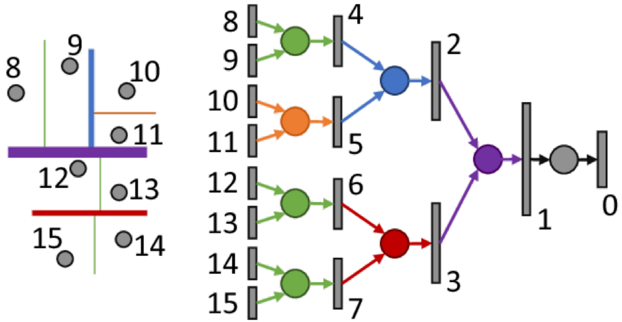
PointNet



PointNet [Qi et al. CVPR 2017]

- Feed the network directly with points and without prior transformation.
- Takes n points (x, y, z) as input.
- Applies input and feature transformations.
- The output is classification scores for k classes.

Kd-Networks



Kd-Network with N=15 points
[Klokov et al. ICCV 2017]

- A kd-tree of depth D is produced with $N = 2^{D-1}$ non-leaf nodes.
- The output is classification scores for k classes.

R-CovNet

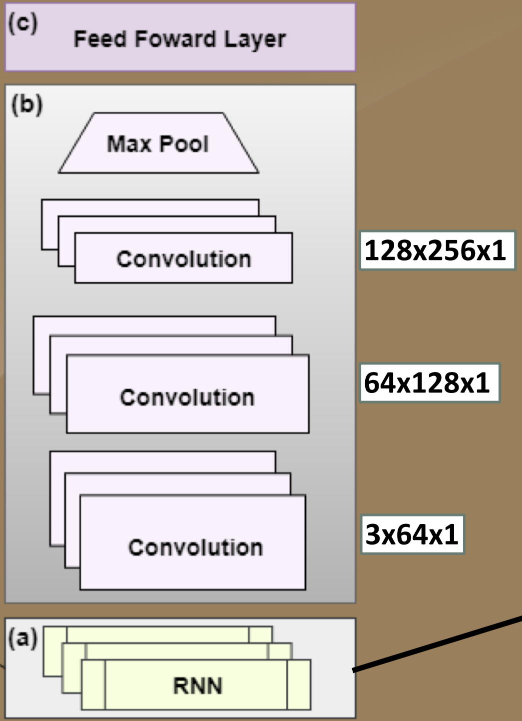
1. **Input**: point cloud with optional additional data (color) as a 1D sequence with 3 channels (x, y, z).

2. **Issues to solve**:
    1. Permutation Invariant
    2. Handling Very Long Sequences
    3. Point clouds are not a time sequence: be able to learn higher order features

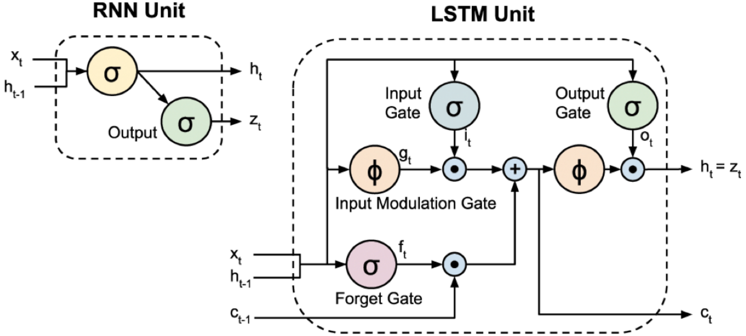
R-CovNet

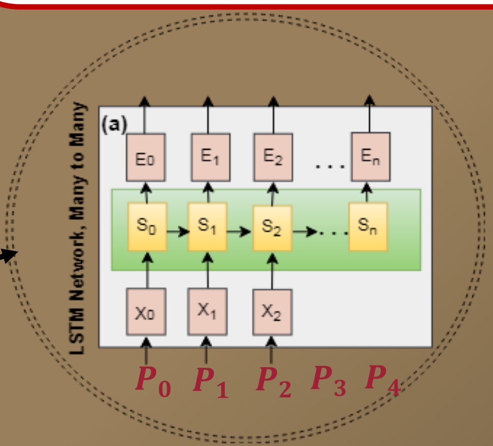
1. Permutation Invariant:
   - Produces the same output regardless of the order of elements in the input
   - RNN is invariant to permutation works well with time sequence.
2. Handling Very Long Sequences
   - Gradient vanishing using classic RNN over time
   - Use LSTM
3. Point clouds are not a time sequence: CNN to learn higher order features.

R-CovNet

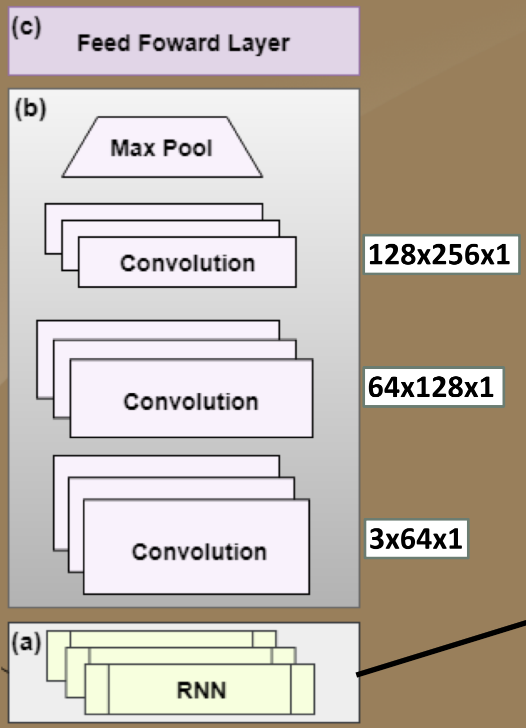


(a) LSTM extension of RNN with higher efficient memory:





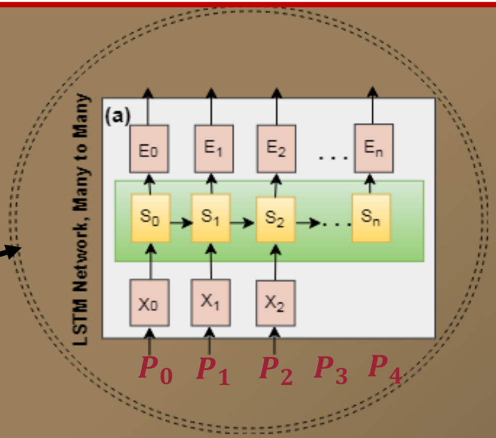
R-CovNet



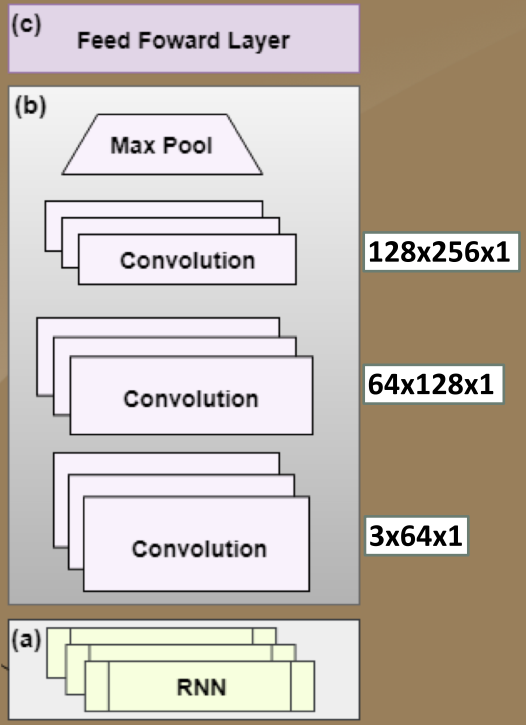
(b) Each layer combination:

- 1D Convolution over RNN output composed of several input planes.
- Batch normalization
- RLU

(c) FC + SoftMax activation: $P(y = j|X) \;=\; \log(\frac{1}{ae_i^x})$; $a = \sum_{j=1}^{n} x^j$



R-CovNet



- (b) Each layer is a combination:
  - Convolution,
  - Batch normalization
  - RLU
- (c) FC + SoftMax activation: $P(y = j|X) = \log(\frac{1}{ae_i^x})$;with a = $\sum_{j=1}^{n} x^j$

Training:

- (a): Backpropagation Through Time (BTT):
  - Data are sent following a time step
  - Gradient depends on the current time step and the previous one.
  - Once the RNN is unfolded, the procedure is analogue to the standard backpropagation
- (b) and (c): Stochastic Gradient Descent (SGD)

Experiments and Results

Implementation details

- Momentum: 0.9
- Batch size:16
- Dataset:
  - ModelNet10: 4899 CAD in 10 classes
  - ModelNet40: 12311 CAD in 40 classes
- Data augmentation: 3D rotation and translation

- GeForce 9300 GE GPU
- ModelNet10: GRU
- ModelNet40: LSTM
- Learning rate varying from 0.01 to 0.00001

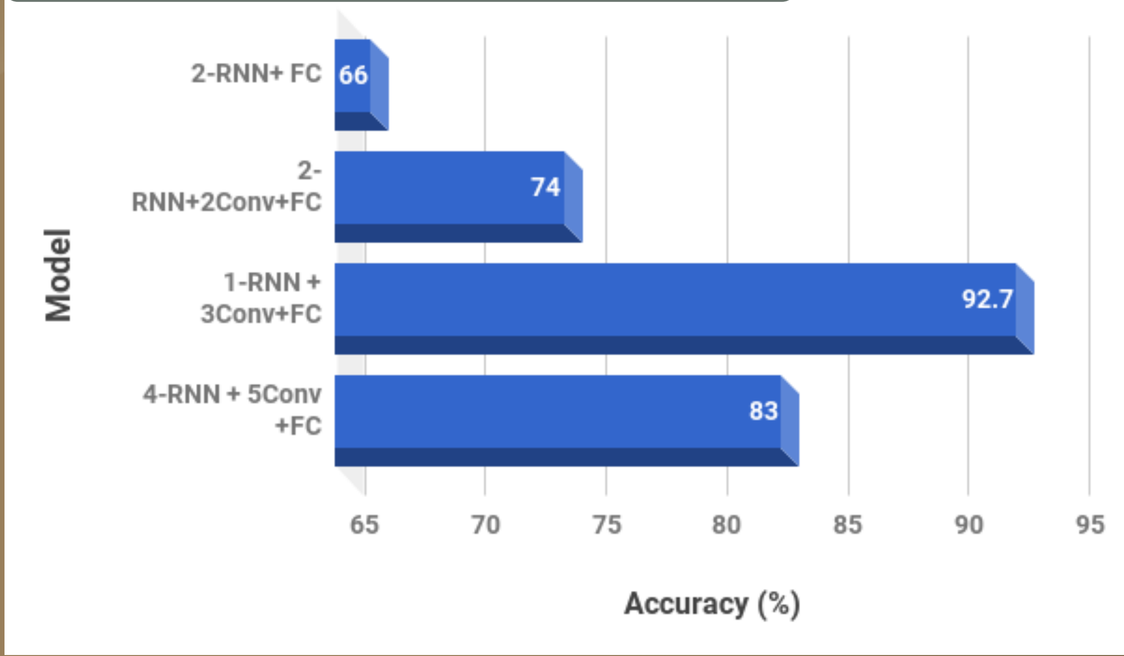
Experiments and Results

Performance

Method	Input	ModelNet10	ModelNet40
VoxNet[2]	Volumetric	92.0%	85.9%
PointNet[3]	Point	-	89.2%
3D ShapeNets[12]	Volumetric	83.54%	77.32%
Kd-networks[4]	Point	93.3%	90.6%
DeepPano[17]	Point	88.66%	82.54%
Set-Conv[18]	Volumetric	-	90.0%
R-ConvNet	Point	**92.7%**	**90.1%**

Experiments and Results

Evaluation on different architectures



Conclusion

- R-CovNet is a novel deep learning approach that process point clouds of different size
- Invariant to permutation.
- Robust to long input sequence
- Made of a combination of RNN and CNN
- Achieve competitive results compare to the current state-of-the art benchmarks

References

[2]Daniel Maturana and Sebastian Scherer, “Voxnet: A 3d convolutional neural network for real-time object recognition,” in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015, pp. 922–928.

[3] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” arXiv preprint arXiv:1612.00593, 2016.

[4] Roman Klokov and Victor Lempitsky, “Escape from cells: Deep kd-networks for the recognition of 3d point cloud models,” arXiv preprint arXiv:1704.01222, 2017.

[12] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao, “3d shapenets: A deep representation for volumetric shapes,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.

[17] Baoguang Shi, Song Bai, Zhichao Zhou, and Xiang Bai, “Deeppano: Deep panoramic representation for 3d shape recognition,” IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2339–2343, 2015.

[18] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos, “Deep learning with sets and point clouds,” arXiv preprint arXiv:1611.04500, 2016.

Thank You