DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks

Yu Xiang and Dieter Fox
Paul G. Allen School of Computer Science & Engineering
University of Washington

ROBOTICS SCIENCE AND SYSTEMS

❏ 3D Scene Understanding

- Navigation
- Manipulation
- ...

- Geometry
  - ✓ Free space
  - ✓ Surface

- Semantics
  - ✓ Objects
  - ✓ Affordances

❏ Related Work

- 3D Scene Reconstruction

  - Newcombe et al., ISMAR'11
  - Henry et al., IJRR'12, 3DV'13
  - Whelan et al., RSSW'12, RSS'15
  - Keller et al., 3DV'13

  ✓ Geometry
  ✓ Data Association
  ✗ Semantics

  KinectFusion

- Semantic Labeling

  - Long et al., CVPR'12
  - Zheng et al., ICCV'15
  - Chen et al., ICLR'15
  - Badrinarayanan et al., CVPR'15

  ✗ Geometry
  ✗ Data Association
  ✓ Semantics

- Semantic Mapping
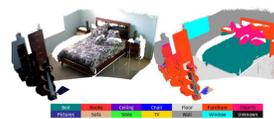
  - Salas-Moreno et al., CVPR'13
  - McCormac et al., ICRA'17

  ✓ Geometry
  ✓ Data Association
  ✓ Semantics

  SemanticFusion

❏ Our Contribution

Recurrent Neural Network

RGB Images → Data Association → Semantic Labels

Depth Images → KinectFusion → 3D Semantic Scene
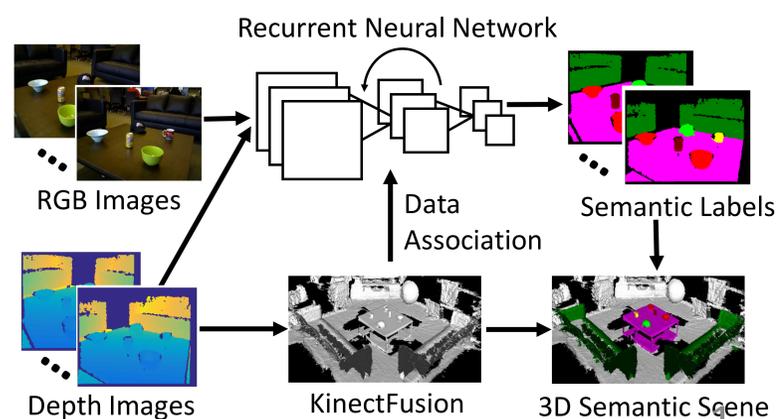
❏ Single Frame Labeling with FCN

Feature Extraction | Embedding | Classification

RGB Image — 64 128 256 512 512 1024 64 — Labels
Depth Image — 64 128 256 512 512 1024 64 #classes

- Convolution + ReLU
- Max Pooling
- Concatenation
- Deconvolution
- Addition
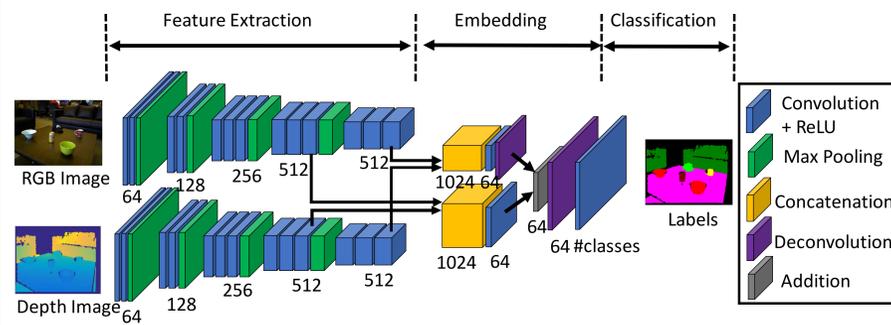
❏ Video Semantic Labeling with DA-RNN
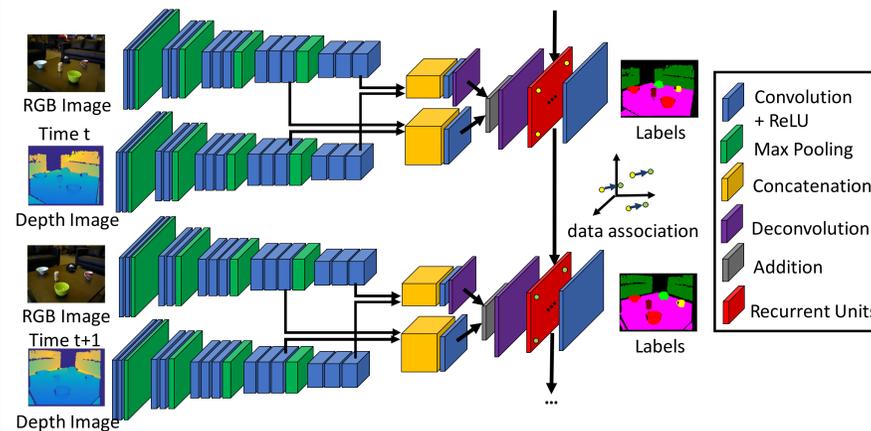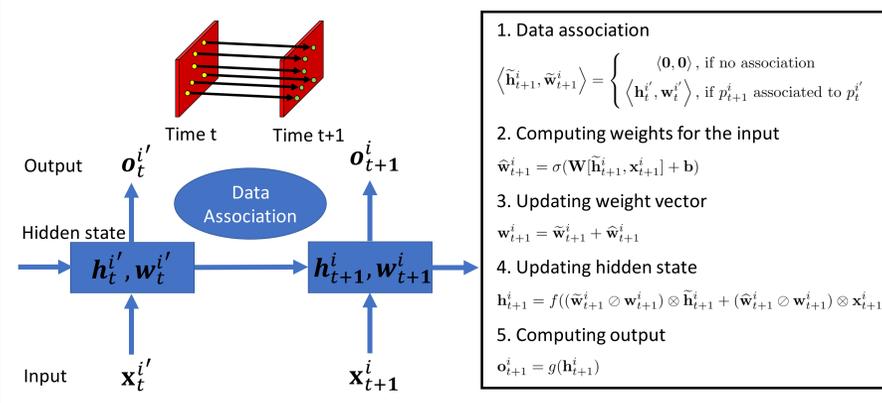
RGB Image Time t
Depth Image

RGB Image Time t+1
Depth Image

data association

Labels

- Convolution + ReLU
- Max Pooling
- Concatenation
- Deconvolution
- Addition
- Recurrent Units

- Data Associated Recurrent Units

Output $o_t^{i'}$ ... Time t ... Time t+1 ... o_{t+1}^i

Data Association

Hidden state $h_t^{i'}, w_t^{i'}$ ... h_{t+1}^i, w_{t+1}^i

Input $\mathbf{x}_t^{i'}$ ... $\mathbf{x}_{t+1}^i$

1. Data association

$$\langle \widetilde{\mathbf{h}}_{t+1}^i, \widetilde{\mathbf{w}}_{t+1}^i \rangle = \begin{cases} \langle \mathbf{0}, \mathbf{0} \rangle, & \text{if no association} \\ \langle \mathbf{h}_t^{i'}, \mathbf{w}_t^{i'} \rangle, & \text{if } p_{t+1}^i \text{ associated to } p_t^{i'} \end{cases}$$

2. Computing weights for the input

$$\widehat{\mathbf{w}}_{t+1}^i = \sigma(\mathbf{W}[\widetilde{\mathbf{h}}_{t+1}^i, \mathbf{x}_{t+1}^i] + \mathbf{b})$$

3. Updating weight vector

$$\mathbf{w}_{t+1}^i = \widetilde{\mathbf{w}}_{t+1}^i + \widehat{\mathbf{w}}_{t+1}^i$$

4. Updating hidden state

$$\mathbf{h}_{t+1}^i = f((\widetilde{\mathbf{w}}_{t+1}^i \oslash \mathbf{w}_{t+1}^i) \otimes \widetilde{\mathbf{h}}_{t+1}^i + (\widehat{\mathbf{w}}_{t+1}^i \oslash \mathbf{w}_{t+1}^i) \otimes \mathbf{x}_{t+1}^i)$$

5. Computing output

$$\mathbf{o}_{t+1}^i = g(\mathbf{h}_{t+1}^i)$$

❏ Experiments

- Comparison on Network Architectures

Methods	FCN [1]	Our FCN	Our GRU-RNN	Our DA-RNN	No Data Association
Background	94.3	96.1	96.8	**97.6**	69.1
Bowl	78.6	87.0	86.4	**92.7**	3.6
Cap	61.2	79.0	82.0	**84.4**	9.9
Cereal Box	80.4	87.5	87.5	**88.3**	14.0
Coffee Mug	62.7	75.7	76.1	**86.3**	4.5
Coffee Table	93.6	95.2	96.0	**97.3**	68.0
Office Chair	67.3	71.6	72.7	**77.0**	13.6
Soda Can	73.5	82.9	81.9	**88.7**	5.9
Sofa	90.8	92.9	93.5	**95.6**	35.6
Table	84.2	89.8	90.8	**92.8**	20.1
MEAN	78.7	85.8	86.4	**90.1**	24.4

Metric: segmentation intersection over union (IoU)

- Analysis on Network Inputs

RGB-D Scene Dataset [2] — IoU vs Inputs (RGB, Depth, Normal, RGB-D)
ShapeNet Scene Dataset [3] — IoU vs Inputs (RGB, Depth, Normal, RGB-D)

FCN / DA-RNN
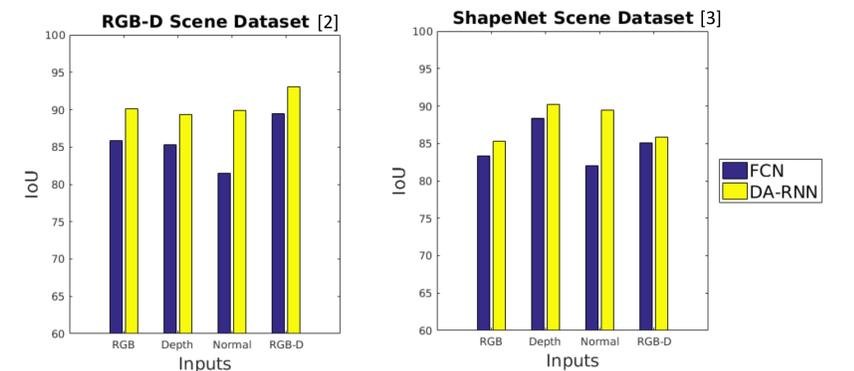
[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR'15.
[2] K. Lai, L. Bo and D. Fox. Unsupervised feature learning for 3D scene labeling. In ICRA'14.
[3] Chang et al., ShapeNet: an information-rich 3D model repository. arXiv preprint arXiv:1512.03012, 2015.

RGB-D Scenes

ShapeNet Scenes

RGB Image | Depth Image | Our FCN | Our DA-RNN | Our DA-RNN 3D