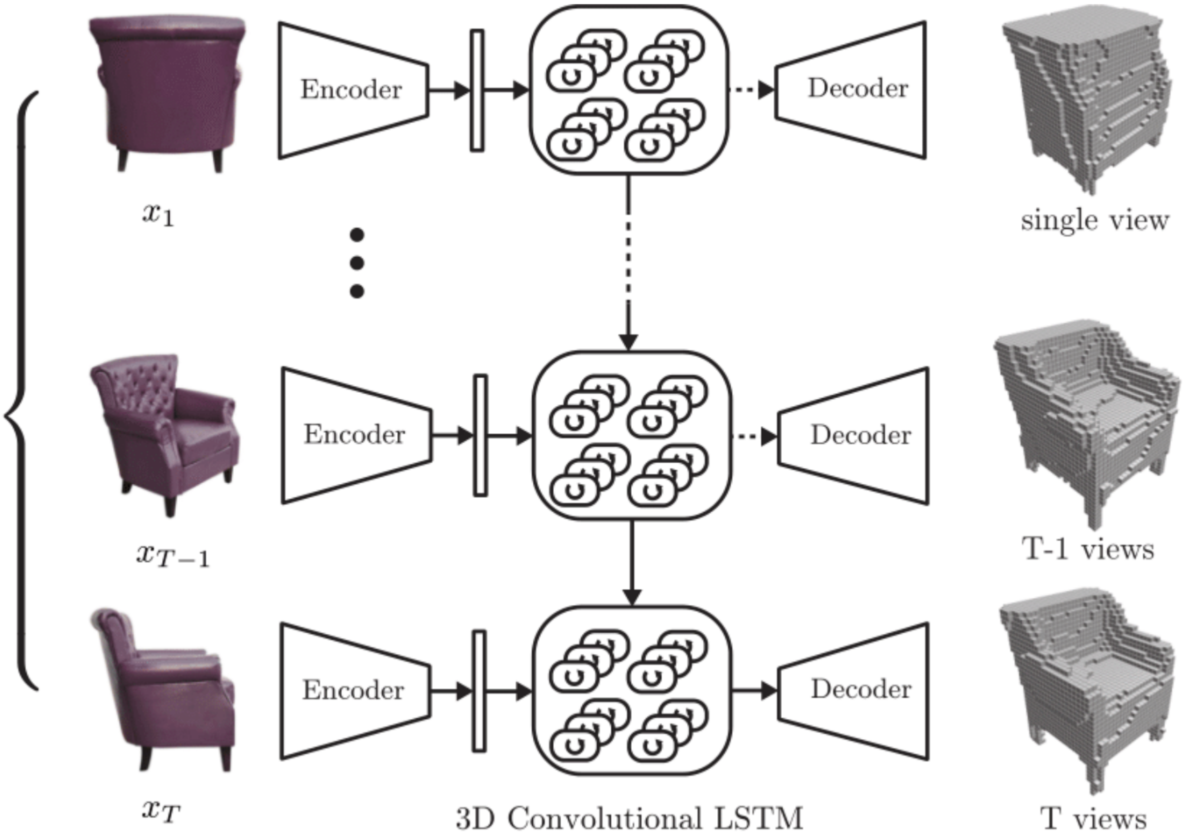
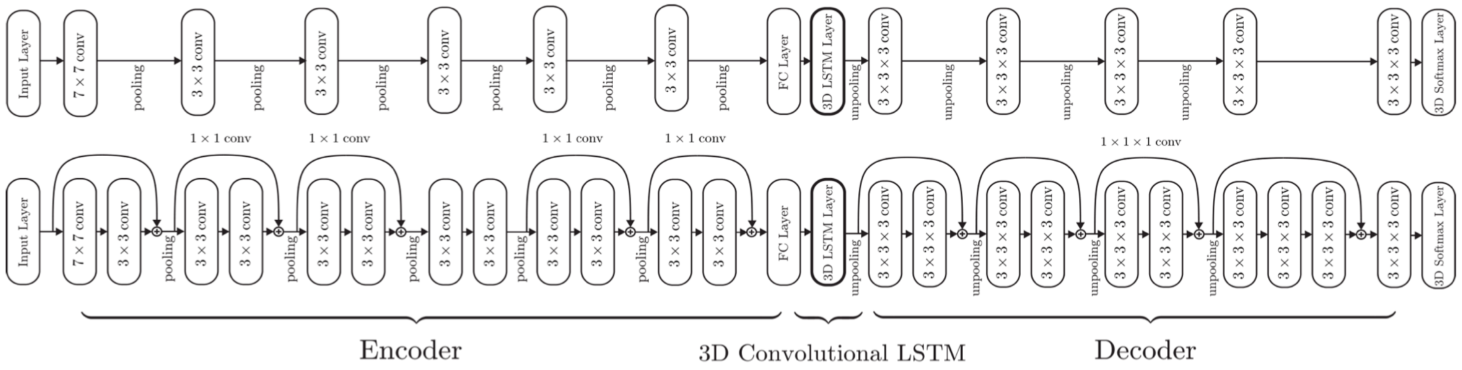
3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction

Report



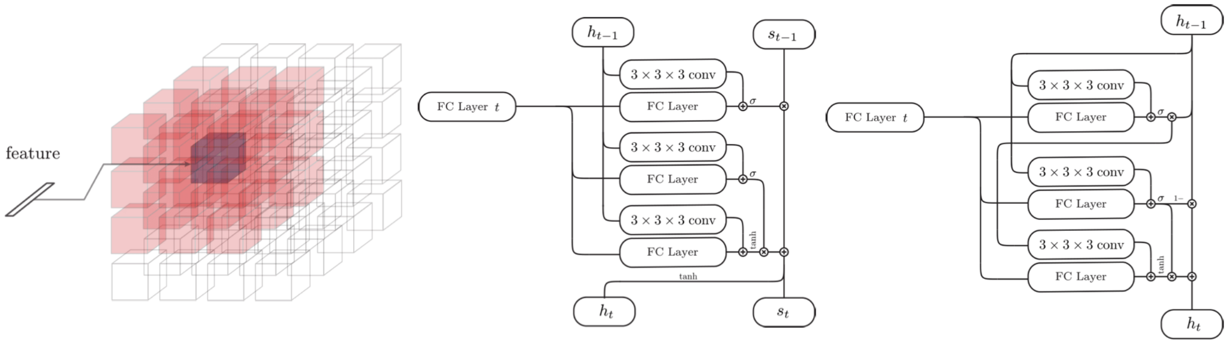
Dataset used: Shapenet, PASCAL 3D
Pytorch Code: https://github.com/chrischoy/3D-R2N2
Summary:
https://lmb.informatik.uni-freiburg.de/lectures/seminar_brox/seminar_ss16/3DR2N2_3Dreconstruction.pdf
Paper: https://arxiv.org/pdf/1604.00449.pdf

Key Points to note:



- They have used LSTM, which is useful for storing information when we use Multi-view images for reconstruction.
- Basically, we feed the information one image at a time, so if we want to reconstruct using 3 orthogonal projections. After feeding the first image, now when we feed the next image, then basically the LSTM remembers the previous image and tries to reconstruct using the previous information
- So in training when we feed in a new image, then the LSTM forgets the previous information and retains only the important information(generic to all images)
- We make use of Residual Connections like Resnet to speed up optimization

3D-LSTM



(a) inputs for each LSTM unit (b) 3D Convolutional LSTMs (c) 3D Convolutional GRUs

- The Fully connected is passed on to the LSTM. This is a 3D LSTM. In a vanilla-LSTM, all elements in the hidden layer ht-1 affect the current hidden state h□ but in 3D LSTM only the neighbouring LSTM affect the hidden state.
- Basically, ht-1 is 3x3x3 space since it receives the previous state from the neighbouring LSTM and we apply 3x3x3 conv to convert it to a proper dimension
- Each unit in the LSTM learns to reconstruct only a part of the voxel space and not the entire space. This also gives a sense of locality so it can selectively update its prediction.
- We can also use GRU instead of LSTM and results have shown that GRU performs much better than LSTM

Decoder

- After receiving an input image sequence x1, x2, · · ·, xT, the 3D-LSTM passes the hidden state h□ to a decoder, which increases the hidden state resolution by applying 3D convolutions, non-linearities, and 3D unpooling until it reaches the target output resolution.