Name: Jiayu Wang
BU email: jiayuw@bu.edu

Project 1, Topic 9: From Images to 3D point clouds

Problem Statement:

Three-dimensional (3D) representations of real-life objects are a core tool for vision, robotics, medicine, augmented reality and virtual reality applications. Point clouds are becoming increasingly popular as a homogeneous, expressive and compact representation geometry, with the ability to represent geometric details while taking up little space. In this topic, we will focus on how to improve quality and speed of generation of 3D point clouds.

Briefly, we will take a 2D image as input into a system, and the output is 3D point clouds, which is process of generating a 3D model. From the perspective of generating images, I think generative adversarial network (GAN) is a good approach for this topic.

Application

3D point clouds have many applications, the most used is iPhone's face ID. Moreover, it has been widely used in construction industry, and interior design industry. Before we have this technical tool, designers needed a lot of time to model the house by drawing. What is even more inconvenient is that it is difficult for customers to visualize and understand these drawings. Now they just need to take an iPad and go around the house, which is much more convenient, Intuitive and specific. This technology, 3D point clouds, is also used in robotics, medicine and other industry fields.

Literature Review

There are several popular frameworks of deep generative models, including generative adversarial networks (GAN), variational auto-encoders (VAE), auto-regressive models, and flow-based models. Flow-based models and auto-regressive models can both perform exact likelihood evaluation, while flow-based models are much more efficient to sample from. I am going to introduce how to use VAE to generate 3D point clouds.

In paper [7], they present multiresolution tree-structured networks to process point clouds for 3D shape understanding and generation tasks. Their model also allows unsupervised learning of point-cloud based shapes by using a variational autoencoder, leading to higher-quality generated shapes. The main contribution of their work is a multiresolution tree network capable of both recognizing and generating 3D shapes directly as point clouds. An overview of the network and how it can be applied to different applications are shown in Figure 1.

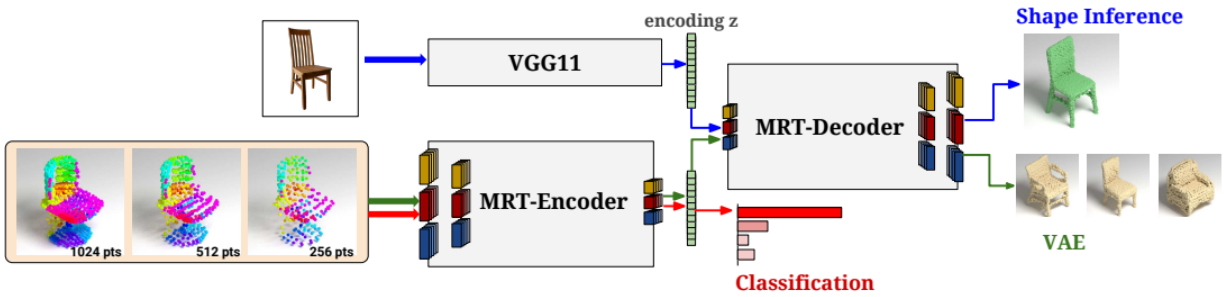

Figure 1 Overview of the network and its applications

Overview of MRTNet. On the left, the MRT-Encoder takes as input a 1D ordered list of points and represents it at multiple resolutions. Points are colored by their indices in the list. On the right, the MRT-Decoder directly outputs a point cloud. Our network can be used for several shape processing tasks, including classification (red), image-to-shape inference (blue), and unsupervised shape learning (green).

In this paper, they represent 3D shapes as a set of locality-preserving 1D ordered list of points at multiple resolution levels. Their multiresolution decoders can be used for directly generating point clouds. This allows us to incorporate order-invariant loss functions, such as Chamfer distance, over point clouds during training. Moreover, it can be plugged in with existing image-based encoders for image-to-shape inference tasks. Their method can both preserve the overall shape structure as well as fine details.

Refer to Fig. 2 for details on the encoder and decoder.

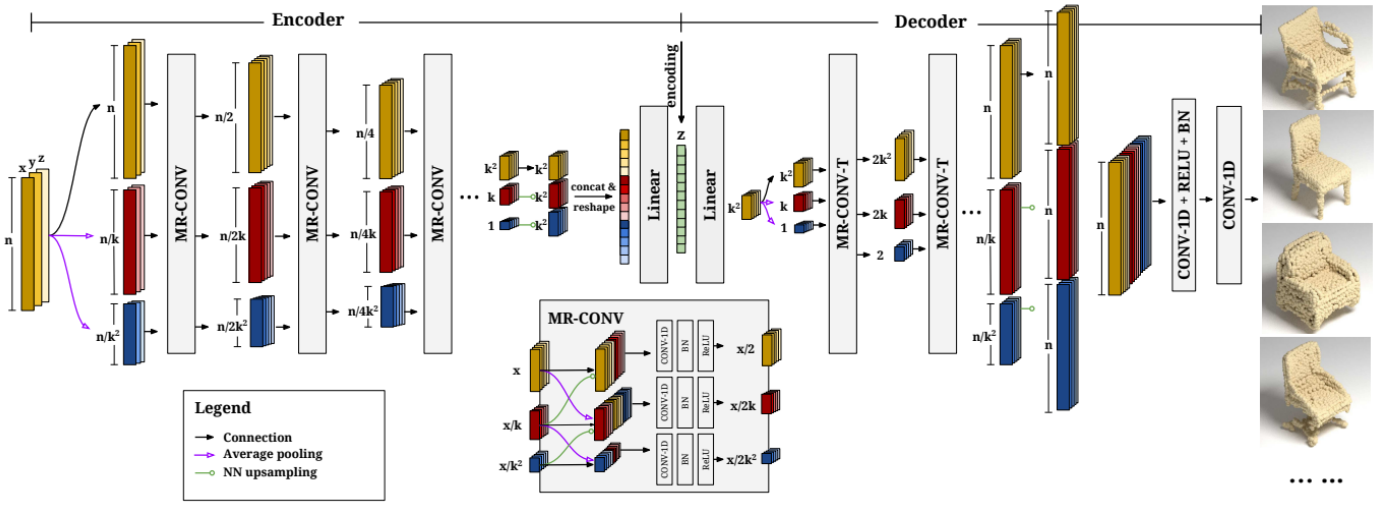

Figure2. Encoder and Decoder

Their multiresolution tree network (MRTNet) for processing 3D point clouds. We represent a 3D shape as a 1D list of spatially sorted point cloud. The network represents each layer at three scales (indicated by yellow, red, and blue), the scale ratio is k between each two. The last two convolution layers have kernel size 1 and stride 1. MR-CONV refers to multi-resolution convolution (zoom-in to the inset for details); and MR-CONV-T refers to MR-CONV with transposed convolution. Our network is flexible and can be used for several shape processing tasks.

shows the complete architecture of our multiresolution tree network (MRTNet) that includes both the encoder and decoder. We represent 3D shapes as a point cloud of a fixed size $N = 2^D$ (e.g., N = 1K). We center the point cloud at the origin and normalize its bounding box; then spatially sort it using a space-partitioning tree. The input to the network is thus a 1D list (N×3 tensor) containing the xyz coordinates of the points. The network leverages 1D convolution and represents each layer at three scales, with a ratio of k between each two. MR-CONV refers to multi-resolution convolution, and MR-CONV-T refers to MR-CONV with transposed convolution. The encoding z is a 512-D vector. Their network architecture is flexible and can be used for several shape processing tasks. For unsupervised learning of point clouds, they use both the multiresolution encoder and decoder, forming a variational autoencoder.

By combining the multiresolution encoder and decoder together, we can perform unsupervised learning of 3D shapes. The entire network, called MR-

VAE, builds upon a variational autoencoder (VAE) [23] framework. The encoder Q receives as an input a point cloud x and outputs an encoding $z \in R^{512}$. The decoder D tries to replicate the point cloud x from z. Both encoder and decoder are built using a sequence of MR-CONV blocks as in Fig.2.

They use Chamfer distance as the reconstruction loss function. Besides this, we also need a regularization term that forces the distribution of the encoding z to be as close as possible to the Gaussian N (0, I). Differently from the original VAE model, we found that we can get more stable training if we try to match the first two moments (mean and variance) of z to N (0, I). Mathematically, this regularization term is defined as:

$$\mathcal{L}_{reg} = \|cov(Q(\mathbf{x}) + \delta) - I\|_2 + E[Q(\mathbf{x}) + \delta]$$

where cov is the covariance matrix, Q is the encoder, k·k2 is the Frobenius norm and E[·] is the expected value. δ is a random value sampled from N (0, cI) and c = 0.01. Thus, our generative model is trained by minimizing the following loss function:

$$\mathcal{L} = Ch(\mathbf{x}, D(Q(\mathbf{x}))) + \lambda L_{reg}$$

Set λ = 0.1. The model is trained using an Adam optimizer with learning rate 10^{-4} and β = 0.9.

For unsupervised learning of point clouds, we train our MR-VAE using the ShapeNet dataset [6]. By default, we compute N = 4K points for each shape using Poisson Disk sampling [3] to evenly disperse the points. Each point set is then spatially sorted using a kd-tree. Here we use the vanilla kd-tree where the splitting axes alternate between x, y, z at each level of the tree. The spatially sorted points are used as input to train the MR-VAE network (Section 3). Like before, we also train a baseline model that follows the same network but replacing multiresolution convolutions with singlescale 1D convolutions in both encoder and decoder. As Figure 3 shows, the shapes generated by the MR-VAE trained on chairs are of considerably higher quality than those generated by the baseline model. They also performed multiple-category shape generation by training MR-VAE on 80% of the objects from ShapeNet dataset. The remaining models belong to our test.

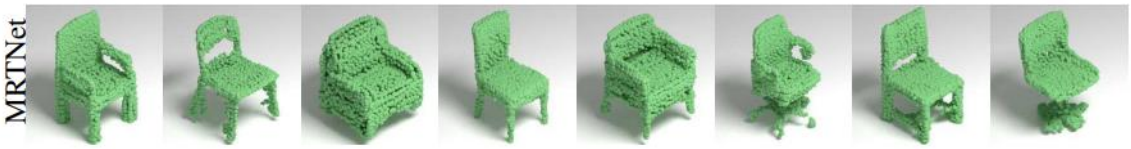

Figure3 Result of 3D point clouds generation

Results are generated by randomly sampling the encoding z. MR-VAE is able to preserve shape details much better than the baseline model and produces less noisy outputs.

Initial Material to Study

1. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L. and Bennamoun, M., 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *43*(12), pp.4338-4364.
2. Fan, H., Su, H. and Guibas, L.J., 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 605-613).
3. Achlioptas, P., Diamanti, O., Mitliagkas, I. and Guibas, L., 2018, July. Learning representations and generative models for 3d point clouds. In *International conference on machine learning* (pp. 40-49). PMLR.
4. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S. and Hariharan, B., 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4541-4550).
5. Lin, C.H., Kong, C. and Lucey, S., 2018, April. Learning efficient point cloud generation for dense 3d object reconstruction. In *proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
6. https://github.com/optas/latent_3d_points
7. Gadelha, M., Wang, R. and Maji, S., 2018. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 103-118).