Unsupervised Feature Learning for Point Cloud by Contrasting and Clustering With Graph Convolutional Neural Network

Ling Zhang and Zhigang Zhu
The City College of The City University of New York
lzhang006@citymail.cuny.edu zzhu@ccny.cuny.edu

Abstract

To alleviate the cost of collecting and annotating large-scale point cloud datasets for 3D scene understanding tasks, we propose an unsupervised learning approach to learn features from unlabeled point cloud "3D object" dataset by using part contrasting and object clustering with deep graph neural networks (GNNs). In the contrast learning step, all the samples in the 3D object dataset are cut into two parts and put into a "part" dataset. Then a contrast learning GNN (ContrastNet) is trained to verify whether two randomly sampled parts from the part dataset belong to the same object. In the cluster learning step, the trained ContrastNet is applied to all the samples in the original 3D object dataset to extract features, which are used to group the samples into clusters. Then another GNN for clustering learning (ClusterNet) is trained to predict the cluster IDs of all the training samples. The contrasting learning forces the ContrastNet to learn high-level semantic features of objects but probably ignores low-level features, while the ClusterNet improves the quality of learned features by being trained to discover objects that belong to the same semantic categories by using cluster IDs. We have conducted extensive experiments to evaluate the proposed framework on point cloud classification tasks. The proposed unsupervised learning approach obtained comparable performance to the state-of-the-art unsupervised learning methods that used much more complicated network structures. The code and an extended version of this work is publicly available via: https://github.com/lingzhang1/ContrastNet

1. Introduction

With ever increasing applications, 3D scene understanding with deep graph convolution neural networks (GNNs) has drawn extensive attention [12, 13, 15, 3]. GNNs typically have millions of parameters which could easily lead to over-fitting. Large-scale annotated datasets are needed for the training of such deep networks. However, the collection and annotation of point cloud datasets are very time-consuming and expensive since pixel-level annotations are needed. With the powerful ability to learn useful representations from unlabeled data, unsupervised learning methods, sometimes also known as self-supervised learning methods, have drawn significant attention.

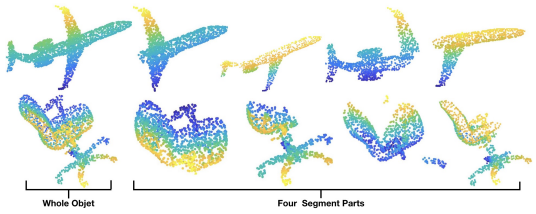

Figure 1. Each row consists of a 3D point cloud object and its four different segments. Human can easily recognize the objects and the locations of the segments in the objects even they are small segments. Inspired by this observation, we propose to train GNNs to learn features from unlabeled dataset by recognize whether two segments are from the same object.

The general pipeline of unsupervised learning with a deep neural network is to design a "pretext" task for the network to solve while the label for this pretext tasks can be automatically generated based on the attributes of the data. After the network is trained with pretext tasks, the network will be able to capture useful features. Recently, many unsupervised learning methods have been proposed to learn image features by training networks to solve pretext tasks [10, 2, 7, 11, 8].

Some unsupervised learning methods have also been proposed for point cloud unsupervised learning [9, 4, 5, 16, 1, 17]. Most of them are based on Auto-Encoder (AE) [5, 16, 1, 17]. Various AEs are proposed and the features are obtained by training AEs to reconstruct the 3D point cloud data. Since the main purpose of an AE is to reconstruct the data, the networks may memorize the low-level features of the point cloud.

In this paper, we propose an unsupervised feature learning approach for point cloud by training GNNs to solve two pretext tasks consecutively, which are part contrasting

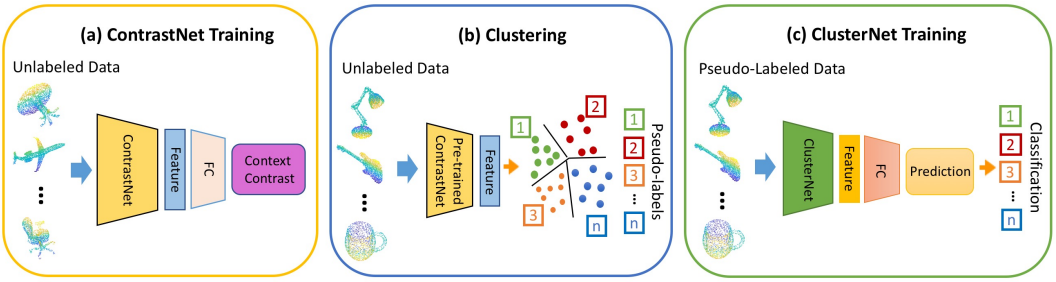

Figure 2. The proposed unsupervised feature learning includes three main steps: (a) ContrastNet for part contrast learning, by verifying whether two point cloud cuts belong to the same object; (b) Clustering samples of 3D objects and assign cluster IDs, using the features learned by ContrastNet; (c) ClusterNet for object clustering learning, by training the network with the 3D point cloud data while the labels are the cluster IDs assigned by the clustering step.

and object clustering. Specifically, the network is trained to accomplish two pretext tasks: to compare (contrast) two point cloud cuts and to cluster point cloud objects. First, all the 3D point objects are cut into two parts and a GNN (called ContrastNet) is trained to verify whether two randomly sampled parts from the dataset belong to the same object. Second, the point cloud data is clustered into clusters by using the features learned by the ContrastNet, and another GNN (called ClusterNet) is trained to predict the cluster ID of each point cloud data. The contrasting learning forces the ContrastNet to learn high-level semantic features while ignoring low-level features, and the predicted cluster IDs boost the quality of learned features by training the ClusterNet to discover objects that belong to the same semantic categories.

In summary, our main contributions in this paper are as follows:

- A generalized and effective unsupervised feature learning framework is proposed for point cloud data. By training deep neural networks to solve two pretext tasks, part contrasting and object clustering, the networks are able to learn semantic features for point cloud data without using any annotations.

- In particular, aligning pseudo-labels for point clouds using clustering is able to transfer knowledge from pre-training models to fine-tuning models. This step significantly boosts the classification performance, a 2.9% improvement on ModelNet40.

- The extensive experiments show that our proposed approach outperforms most of the state-of-the-art unsupervised learning methods. With the features learned from the unlabeled dataset, the proposed model obtains 86.8% and 93.8% on ModelNet40 and ModelNet10 dataset respectively.

2. Method

To learn features from unlabeled point cloud data, we propose to learn features by training networks to accomplish both of the part contrasting and the object clustering pretext tasks. The pipeline of our framework is illustrated in Fig. 2, which includes three major steps: ContrastNet for part contrast learning, clustering using the learned features, and then ClusterNet for object cluster learning using the cluster IDs.

2.1. ContrastNet: Part Contrast Learning

When a point cloud data is observed from different views, only a part of the 3D object can be seen. The observable part can be very different based on the view, as shown in Fig. 1. Inspired by this observation, we propose to use part (segment) contrasting as a pretext task for a GNN to solve. The task is defined as to train a ContrastNet to verify whether two point cloud segments belong to the same object. The positive pair is drawn by selecting two different segments from the same object, while the negative pair is drawn by selecting two segments from two different objects. The illustration of the part contrast task is shown in Fig. 3.

This task can be modeled as a binary classification problem. As the training goes on, the segments from the same object should have a smaller distance while the segments from different objects have a larger distance. In this way, semantic features can be learned by this process. DGCNN [15] is used as the backbone model and the details of the network architecture are shown in Fig. 3.

2.2. ClusterNet: Knowledge Transfer with Clusters

The underline intuition of clustering is that 3D objects from the same categories have high similarity than those from different categories. After obtaining the clusters of the data by using Kmeans++[6], based on the features extracted by ContrastNet, the cluster IDs of the data are used as the ”pseudo” labels to train a ClusterNet. We hope that

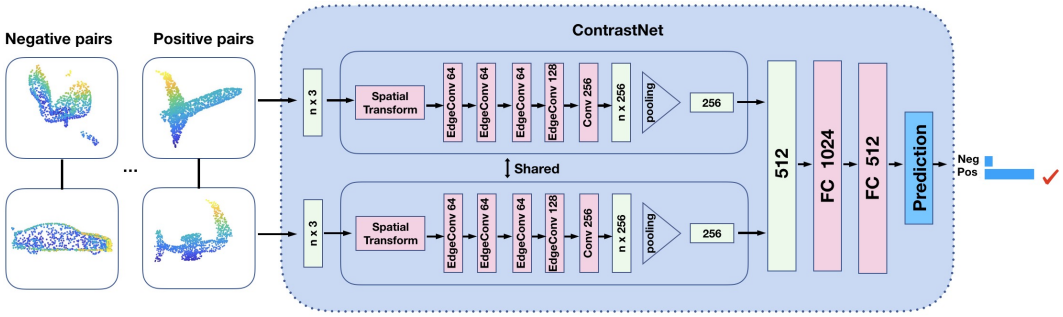

Figure 3. The architecture of ContrastNet for part contrast learning. The positive pair is generated by randomly sampling two segments from the same point cloud sample, while the negative pair is generated by randomly sampling two segments from two different samples. A dynamic graph convolution neural network (GNN) is used as the backbone network. The features of two segments are concatenated and fed to fully connected layers to make the prediction of positive or negative. The part contrast learning does not require any data annotations by humans.

using cluster IDs as pseudo labels in ClusterNet can provide more powerful self-supervision and therefore, the network can learn more representative features for object classification.

The training of ClusterNet, also based on DGCNN [15], with the cluster ID assignments as the pseudo-labels, is described as:

$$\min_{\theta,W} \frac{1}{N} \sum_{n=1}^{N} \ell\left(g_W\left(f_\theta(x_n)\right), y_n\right), \tag{1}$$

where the purpose of training is to find the optimal parameters θ^* such that the mapping f_{θ^*} produces good general-purpose features for point cloud data. In our unsupervised learning training, each data x_n is paired with a pseudo label y_n that is generated by the clustering algorithm.

3. Experimental Results

3.1. Implementation Details

During the part contrast unsupervised learning, each object is cut by 15 randomly generated planes into 30 segments. Each segment has at least 512 points. During the unsupervised part contrast training phase, the learning rate is 0.001, momentum is 0.9, the learning rate decay rate is 0.7, and the decay step is 200000. The same DGCNN structure and the learning parameters are used as in the ClusterNet.

3.2. Transfer Features Learned to Classification Task

To quantitatively evaluate the quality of the learned features by using the part contrasting pretext task and then by the ClusterNet, we conduct experiments on three different datasets: ShapeNet, ModelNet10, and ModelNet40. The features are extracted by the ContrastNet that is only trained with the part contrast task on unlabeled data. A linear classifier SVM is trained based on the features of the training data, and the testing classification accuracy of ContrastNet and ClusterNet are reported in the columns "ContrastNet" and "ClusterNet" in Table 1, respectively.

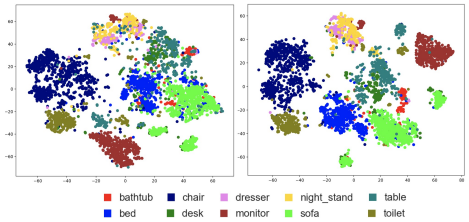

Figure 4. Visualization of object embedding of the ModelNet10 test data through part contrast training on the ShapeNet dataset. The features are learned by part contrast learning (left) and then boosted by object clustering (right).

Training	Testing	ContrastNet (%)	ClusterNet (%)
ShapeNet	ModelNet40	84.1	86.8 (**+2.7**)
ShapeNet	ModelNet10	91.0	93.8 (**+2.8**)
ModelNet40	ModelNet40	85.7	88.6 (**+2.9**)

Table 1. Comparison of 3D object classification results using ContrastNet and ClusterNet. The classification accuracy of ClusterNet have average 2.8% improvement on all experiments.

As shown in Table 1, these results validate the effectiveness of the proposed method and the learned features by the ContrastNet indeed have semantic information. Training the ClusterNet to predict the cluster ID of each data can significantly improve the point cloud classification accuracy at least 2.7% on all three datasets. These improvements validate the effectiveness of using clustering to boost the quality of the learned features, as shown in Fig. 4.

3.3. Compare with the State of the Art

We compare our approach with other unsupervised learning models [9, 4, 5, 14, 16, 1, 17] on point cloud clas-

sification benchmarks ModelNet10 and ModelNet40. Following the common practice [17, 16], all the models are trained on the ShapeNet data with the same procedure. The methods in [9, 4] are hand-crafted features and methods in [5, 14, 16, 1, 17] are deep learning based methods.

On the ModelNet10 dataset, our methods outperforms SPH [9], LFD [4], TLNetwork [5], VConv-DAE [14], and 3DGAN [16], and only 0.6% lower than FoldingNet [17] which is the latest work for unsupervised feature learning. On the MoldelNet40 datasets, our method outperforms all the methods except FoldingNet (1.6% lower). We would like to note that our ClusterNet has a much simpler structure and is easier in training.

Models	ModelNet40 (%)	ModelNet10 (%)
SPH [9]	68.2	79.8
LFD [4]	75.5	79.9
T-L Network [5]	74.4	-
VConv-DAE [14]	75.5	80.5
3D-GAN [16]	83.3	91.0
Latent-GAN [1]	85.7	95.3
FoldingNet [17]	88.4	94.4
ClusterNet (Ours)	**86.8**	**93.8**

Table 2. The comparison on classification accuracy between our ClusterNet and other unsupervised methods on point cloud classificaton dataset ModelNet40 and ModelNet10.

4. Conclusion

We have proposed a straightforward and effective method for learning features for point cloud data from unlabeled data. The experiment results demonstrate that proposed pretext tasks (part contrasting and object clustering) are able to provide essential semantic information of the point cloud data for the network to learn semantic features. Our proposed methods have been evaluated on three public point cloud benchmarks and obtained comparable performance with other state-of-the-art self-supervised learning methods.

5. Acknowledgments

The work is partially supported by the National Science Foundation via awards #CNS-1737533 and #IIP-1827505, as well as a CUNY-Bentley CRA.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Representation learning and adversarial generation of 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2(3):4, 2017. 1, 3, 4

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 1

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[4] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, volume 22, pages 223–232. Wiley Online Library, 2003. 1, 3, 4

[5] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, pages 484–499. Springer, 2016. 1, 3, 4

[6] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. 2

[7] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2018. 1

[8] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint arXiv:1902.06162*, 2019. 1

[9] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003. 1, 3, 4

[10] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 1

[11] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 1

[12] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 1

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017. 1

[14] Abhishek Sharma, Oliver Grau, and Mario Fritz. Vconv-dae: Deep volumetric shape learning without object labels. In *European Conference on Computer Vision*, pages 236–250. Springer, 2016. 3, 4

[15] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018. 1, 2, 3

[16] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, pages 82–90, 2016. 1, 3, 4

[17] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, pages 206–215, 2018. 1, 3, 4