Few-shot 3D Point Cloud Semantic Segmentation
Name: Mayur Mankar
Roll Number: 20169

Abstract

We propose a new method for 3D point cloud semantic segmentation that addresses the limitations of fully supervised approaches. These approaches require large amounts of labeled training data that are difficult to obtain and cannot segment new classes after training. Our approach is attention-aware, multi-prototype, and transductive, allowing us to segment new classes with only a few labeled examples. We use multiple prototypes to represent each class and employ a transductive label propagation method to exploit affinities between labeled multi-prototypes and unlabeled points. Additionally, we use an attention-aware multi-level feature learning network to capture geometric dependencies and semantic correlations between points. Our proposed method shows significant and consistent improvements compared to baselines in different few-shot point cloud semantic segmentation settings on the S3DIS dataset.

Keywords – Point Cloud Semantic Segmentation (PCSS), ProtoNet, Dynamic graph convolutional neural network (DGCNN), Linked Dynamic graph convolutional neural network (LDGCNN).

Introduction

The estimation of each point's category in a 3D point cloud representation of a scene is a challenging computer vision task known as point cloud semantic segmentation. Point clouds are unstructured and unordered, making this task particularly difficult. A variety of fully supervised 3D semantic segmentation techniques have recently been proposed, and they have achieved promising performance on various benchmark datasets. However, their success is heavily reliant on large amounts of labeled training data, which can be time-consuming and costly to collect. Additionally, these methods rely on the closed set assumption, which assumes that the training and testing data are drawn from the same label space. This assumption is not always valid in the dynamic real world, where new classes can emerge after training. As a result, these fully supervised approaches struggle to generalize to new classes with only a few examples.

Existing works have attempted to address the issue of data scarcity in 3D point cloud semantic segmentation through self-, weakly-, and semi-supervised learning techniques. However, these methods still adhere to the closed set assumption and overlook the ability to generalize to new classes. Few-shot learning has gained popularity as a promising approach that enables models to segment new classes using only a few labeled point clouds. This approach uses episodic training, where the model learns over a distribution of similar few-shot tasks instead of a single target segmentation task. Each task comprises a few labeled samples (support set) and unlabeled samples (query set), and the model segments the query set using the knowledge gained from the support set. The model is better able to generalize and avoid overfitting rare support samples due to the consistency between the training and testing few-shot tasks. However, few-shot point cloud segmentation still faces two significant challenges: how to distill discriminative knowledge from scarce support sets that can represent the distributions of novel classes, and how to leverage this knowledge effectively for segmentation.

This paper presents a new approach to few-shot point cloud semantic segmentation that uses attention-aware multi-prototype transductive inference. The proposed method is designed to model the intricate point distributions in the support set and perform segmentation through transductive inference using discriminative features extracted under the few-shot constraint. Inspired by the prototypical network, which represents each class with a single prototype obtained from averaging the embeddings of labeled samples in the support, the authors suggest that the unimodal distribution assumption may be insufficient in point cloud segmentation due to the diverse data distribution of points. As a result, they propose using multiple prototypes to represent each class and better capture the complex distribution of geometric structures within the same semantic class.

In the few-shot 3D point cloud semantic segmentation task, it is crucial to learn features that are discriminative. To achieve this goal, we propose an attention-aware multi-level feature learning network that can capture the semantic correlations and geometric dependencies between points. We then perform the segmentation step using a transductive approach, utilizing multiple prototypes in the learned feature space. In contrast to the traditional prototypical network that matches unlabeled instances with class prototypes based on Euclidean distances, our transductive inference method takes into account both the relationships between the unlabeled query points and the multi-prototypes, as well as the relationships among the query points themselves.

This work presents several significant contributions. Firstly, we investigate the few-shot 3D point cloud semantic segmentation task, which has not been explored before. Secondly, we propose a novel approach that utilizes attention-aware multi-prototype transductive inference. Our method incorporates attention-aware multi-level feature learning, and leverages the affinity between multi-prototypes and unlabeled query points to obtain highly discriminative features and achieve more accurate segmentation in the few-shot scenario. Thirdly, we conduct extensive experiments on the S3DIS dataset and demonstrate the superior performance of our proposed method over existing baselines in various few-shot point cloud segmentation settings.

Dataset

S3DIS dataset

Path to the dataset: https://drive.google.com/file/d/1Wag2wzdLotY8RWBjrc_8u1Rpr7Lqj7NU/view?usp=sharing

The Stanford 3D Indoor Scene Dataset (**S3DIS**) dataset contains 6 large-scale indoor areas with 271 rooms. Each point in the scene point cloud is annotated with one of the 13 semantic categories.

Code

The code is present on the server.

To access the code, follow the steps given below:

1) Log in to the server: ssh -X dl@172.30.1.163
2) Password: dl@iiserb
3) Go to the directory **cd /data4/dl/DL316_22_23_2/grp09/mayur/AI/attMPTI**

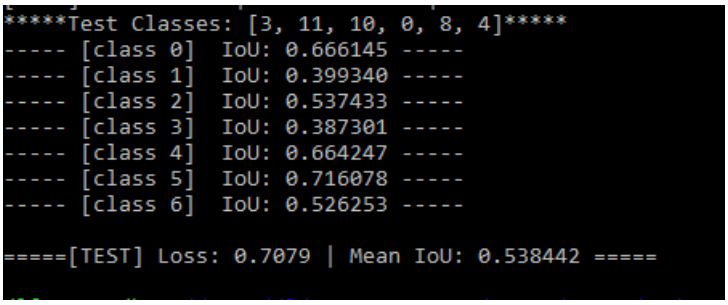
Methodology

1) Installation of required libraries in the virtual environment:
   pip install torch
   pip install faiss-gpu
   pip install tensorboard h5py transforms3d
   pip install pyg_lib torch_scatter torch_sparse torch_cluster torxh_spline conv -f http://data.pyg.org/whl.torch-2.0.0+cu117.html
2) Dowloading and importing the dataset into the working directory:
   a) The S3DIS dataset is downloaded from the following website: http://buildingparser.stanford.edu/dataset.html#Download
   b) Then the dataset is imported from local system into the working directory using scp command.
3) Unzipping the dataset.
4) Re-organizing the raw data into npy files by running following command:
   python preprocess/collect_s3dis_data.py --data_path Stanford3dDataset_v1.2
5) Then splitting rooms into blocks by running following command:
   python preprocess/room2blocks.py --data_path ./datasets/S3DIS/scenes/
6) Pretraining the segmentor which includes feature extractor module on the available training set:
   bash pretrain_segmentor.sh

```
7) Training the method:
      bash scripts/train_attMPTI.sh
8) Evaluating the method:
      bash scripts/eval_attMPTI.sh
```

Results and Discussions

For 2 – way 1 – shot learning, for split = 0, result on **S3DIS** dataset using mean-IoU metric (%). S^i denotes the split *i* is used for testing is as given below:



Implementation Details

1) Pretraining
   Pretraining is done for 100 epochs, where the learning rate is set to 0.001 and the batch size is 16.

2) Training

   Training is done for 2 way 1 shot setting. The number of iterations is set to 10000 with an evaluation interval equal to 500. The learning rate is set to 0.001 and the batch size is set to 32.

3) Evaluation

   Evaluation is done for the method.

Novelty

We use Linked dynamic graph CNN (LDGCNN) in place of dynamic graph CNN (DGCNN) to classify and segment point cloud directly. We remove the transformation network, link hierarchical features from dynamic graphs, freeze the feature extractor, and retrain the classifier to increase the performance of LDGCNN. We optimize the network architecture of DGCNN to increase the performance and decrease the model size of the network. Because our network links the hierarchical features from different dynamic graphs, we call it linked dynamic graph CNN (LDGCNN). The differences between our LDGCNN and DGCNN are as follows:

• We link hierarchical features from different layers.

• We remove the transformation network

References

[1] N. Zhao, T. Chua and G. Lee, "Few-shot 3D Point Cloud Semantic Segmentation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021 pp. 8869-8878. doi: 10.1109/CVPR46437.2021.00876

[2] K. Zhang et al., "Linked Dynamic Graph CNN: Learning through Point Cloud by Linking Hierarchical Features," 2021 27th International Conference on Mechatronics and Machine Vision in Practice (M2VIP), Shanghai, China, 2021, pp. 7-12, doi: 10.1109/M2VIP49856.2021.9665104. *Recognition*.

[3] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, & Silvio Savarese (2016). 3D Semantic Parsing of Large-Scale Indoor Spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern*

[4] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2019. Dynamic Graph CNN for Learning on Point Clouds. ACM Trans. Graph. 1, 1, Article 1 (January 2019), 13 pages. https://doi.org/10.1145/3326362