LGFormer: Local-Global Transformer based Lightweight Point Cloud Semantic Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Point cloud serves as digit representation of three-dimensional scene, its semantic segmentation can provide rich contextual information for high-level understanding. While graph structures inherently describe this information and are adaptable to semantic segmentation tasks, existing graph-based approaches predominantly focus on node features, often overlooking the rich edge information and long-range node relationships, thereby neglecting substantial contextual data present in point clouds. This paper introduces a novel model that integrates graph neural networks with attention mechanisms to comprehensively harness both *local* and *global* contextual features. Basically, our method constructs a point cloud adjacency topology graph derived from superpixel segmentation, followed by the extraction and fusion of local and global contexts through different units. Specifically, local contexts are captured via graph convolutional networks, whereas global contexts are precessed using transformers. Evaluated on the S3DIS dataset, our model achieves a state-of-the-art segmentation accuracy of 90.6%. Moreover, by leveraging superpixels and local-global feature fusion, our model demands considerably fewer computational resources compared to existing models with comparable performance, thus offering an efficient solution for indoor point cloud semantic segmentation tasks. All codes and pre-trained models will be released to the community upon acceptance.

1 Introduction

TODO: NOTE: need to use consistent description of SuperPixel/SuperVoxel/SuperPoint in the paper.

With the rapid advancing of laser scanning technologies, point cloud, as a type of accurate measurement of 3D scenes, has been widely utilized in various industrial applications, such as autonomous driving and robotics [1, 2], where semantic segmentation of point cloud plays an essential role. By obtaining semantic labels from segmentation, computers can accurately comprehend 3D scenes and further accomplish more intricate tasks such as spatial perception and 3D reconstruction.

In recent days, numerous deep learning based studies have been carried out for point cloud semantic segmentation. The current approaches can be generally classified into four categories. Intuitively, due to the irregular and unorganized property of point clouds, how to convert unordered data into an ordered representation will be a very important consideration. Projection-based methods [3] usually transform the point clouds into 2D images by projecting them from various perspectives. After semantically segmenting the corresponding images, the identified labels will be mapped back onto the points. It often yields better results due to the abundance of 2D training data available. However, this method can inevitably lead to the loss of critical information during the conversion, for example, the intrinsic 3D structure of point clouds. In order to better preserve the 3D information, voxel-based segmentation [4] methods is an alternative approach to partition the point clouds into regular voxels, followed by 3D convolutions for downstream feature extraction. This straightforward

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

extension demonstrates an improved performance. However, it brings massive computational and memory cost due to the cubic growth in the size of voxels. To address this issue, point-based methods directly learn to extract the rotational invariant features of each point in high-dimensional space [5, 6]. Although such methods are capcable of learning spatial features effectively, the limited distance of feature aggregation means that these methods can only collect local information to assist semantic understanding. Based on this, PointTransformer [7] introduces the attention mechanism to better learn global feature correlations, thereby significantly enhancing the representation ability of the features, but at the cost of high computational expense. In order to further reduce the redundant computation of features for neighboring points, SuperVoxel or SuperPoint-based methods first cluster the point cloud into supervoxels based on the proximity of nearby points. The features are extracted for each supervoxel, and relations between different supervoxles are depicted via graph neurual networks [8]. Overall, existing methods typically employ complex network architectures to extract more expressive features, but this also introduces additional storage and computational overhead. How to balance these aspects is rarely mentioned in current works, which is the focus of our study.

In this work, we introduce a novel network architecture that integrates graph neural networks with attention mechanisms, to explore the intricate contextual topological information among supervoxels, for enhanced semantic segmentation performance. Our model discriminates the contextual information inherent in point clouds into two distinct categories: local context, which encapsulates the neighborhood features, thereby quantifying the structural information of point cloud; and global context, which contains the relative positioning and inter-object feature relationships within the entire point cloud. To effectively harness these two forms of contexts, our architecture employs two inter-related units. The first unit automatically learns the local topological relationships by examining the adjacency of multiple super voxels, capturing the fine-grained structural details. Concurrently, the second unit adopts a global perspective, augmenting the super voxel features with broader scene-level information to distill comprehensive global features. By exploiting the rich contextual information through the aforementioned architecture, our model achieves superior semantic segmentation results. This cooperative learning strategy ensures a deep understanding of the point cloud, leading to significant advancements in segmentation accuracy and robustness. Note that, our design choice of graph neural network preserves the feature information obtained in the intermediate process to avoid potential information loss from multiple convolution operations, allowing for focus on feature extraction within the current field of view, enhancing the overall learning efficiency of the model and thereby achieving a lightweight network architecture. Overall, TODO: the main contributions of this paper are summarized as follows:

- A lightweight local-global transformer framework for indoor point cloud semantic segmentation.
- A GCN-GRU module which preserves explicit topological relation is proposed to guide the strong constraints to nodes.
- In terms of rebalancing the accuracy and efficiency, our point cloud semantic segmentation achieves SoTA.

2 Related Works

In the following, we will only summarize approaches that are closely related to point cloud semantic segmentation. For a more comprehensive review of deep learning for point clouds and their broader applications, please refer to the survey paper [9–11].

Point cloud feature extraction. The traditional handcrafted feature extraction methods [12, 13] have dominated point cloud semantic segmentation task for years, primarily due to the scarcity of data and the limited range of target classes. However, these methods struggle to cope with complex scenarios, as the number of engineered features is inherently restricted. To this end, deep learning approaches have been introduced to address the challenges. In order to better apply deep learning architecture to unorganized point cloud data, voxel-based methods [14, 4, 15] were initially proposed, aiming to extract voxel-unit features and regularize the data for more effective processing. As we all know, the voxelized point clouds suffer from boundary artifacts and cannot accurately describe boundary information due to the resolution limitation of voxels. Subsequently, point-based methods [5, 6, 16–23] are proposed to mitigate this issue, and directly extract point features for