Research notes on Semantic Segmentation for Point Cloud via Semantic-based Local Aggregation and Multi-scale Global Pyramid

1.1. Research Question

1) How to judge points in the local region belong to the same category?
2) In the junction region with multi-objects of different classes, how to make the local representation meaningful?
3) How to use the global features and identify the muti-scale object with similar spatial structures?
4) How to replace multi-scale receptive fields to extract multi-scale features and solve the problem of increased computational consumption and redundant information?

1.2. Motivation

1) There are three main neural network-based point cloud processing methods: projection-based, voxel-based and point-based. Point-based methods have the advantage that preserve more geometrical details than others. Projection-based methods have the shortage that loss of geometric information, voxel-based networks face the heavy memory consumption and loss of fine-grained geometry information.
2) Most works based on the SA layer have two drawbacks: 1) The SA layer neglects to consider whether the points in the local region belong to the same category. For the junction region with multi-objects of different classes, the local representation may not represent any class. 2) Most works focus on how to extract local features efficiently while ignoring global features (contain more implicit information about large-scale objects than local features, and multi-scale features enable the network to identify objects with similar spatial structures but different scales). So the author want to take full advantage of multi-scale global features to better identify the objects.
3) Because using multi-scale receptive fields will cause increased computational consumption and redundant information, author wants obtain discriminative multi-scale global features without adding multi-scale receptive fields additionally.

1.3. Framework

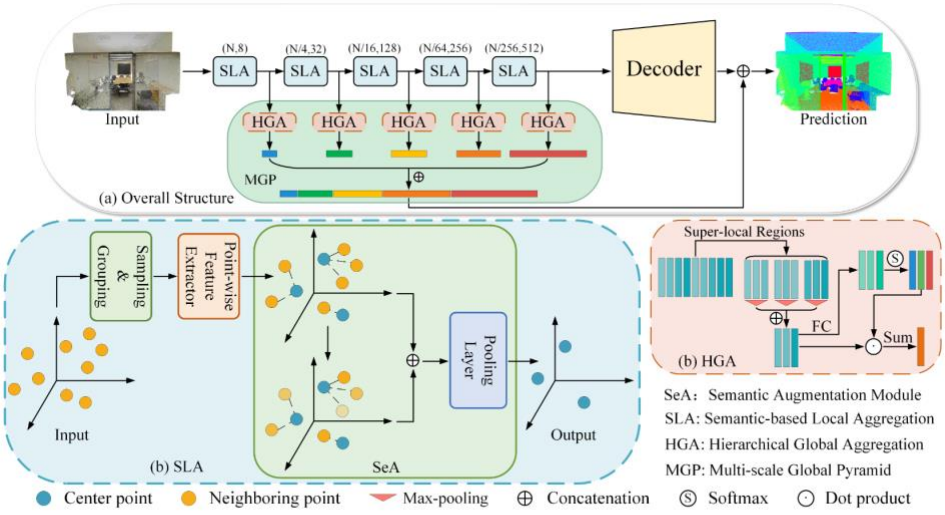

Figure 1. Flowchart of our proposed method. (**a**) is the overall framework. (**b**) is the SLA strategy, we embed a SeA module between the point-wise extractor and the pooling layer to augment local context. (Different transparency indicates different levels of semantic similarity.) (**c**) is the HGA module, and MGP in (**a**) contains five HGAs.

First, the input of the model is the point cloud, and then the local features through semantic similarity will be augmented through the SLA module five times, and the obtained results are fed into the next encoding layer and the HGA module. After each SLA module, an HGA module will be connected. After five HGA modules, the local features at different scales are fused to form a new multi-scale global feature. But here I have a question, why are there 5 SLA modules? If there is one more or one less SLA, how much impact will it have on the computation and obtaining more discriminative multi-scale global features, and whether this number is the result of trade-off.

Moreover, a Semantic Score Block is included in the Semantic Augmentation (SeA) module. Different semantic scores reflect the different levels of semantic similarity possessed by neighboring points. However, due to the limitations of the score function, the author believes that score mapping errors are inevitable. But we can also try to optimize the score function to reduce the negative effects brought by errors. In the Hierarchical Global Aggregation part, the author learned from previous works that attention-based pooling method to aggregate local features achieve better performance than max-pooling, so the author defines global scenes as "super-local regions" and uses attention mechanism to aggregate a global feature from the super-local feature set, so this is a novel hierarchical global aggregation method.

1.4. Experiment

This model based on two benchmarks: S3DIS and Semantic3D. Evaluation on S3DIS shows that this model has more advantages in identifying similar simple spatial structures but different-scale objects, which is in line with the problem that the model originally wanted to solve. Evaluation on Semantic3D shows that this model achieves improvements in high-vegetation and low-vegetation, two classes with similar structures but different scales.

In the Ablation Study section, we can find that if we augmented local context by neighboring features with irrelevant semantics, it reduces the performance, which means that more information will not improve the performance of the model. And error calculations inevitably occur due to insufficient modeling ability, especially in shallow networks.

And we can find that the plug one SeA in the second feature aggregation operation will be the most helpful for performance improvement through the author's experimental comparison. And both SLA and MGP have significantly improved semantic segmentation results.

1.5. The reason of acceptance

1) It is proposed to combine the multi-scale global feature, and does not require additional multi-scale receptive fields, the designed HGP module is very novel.
2) Achieved better performance than previous models in recognizing objects of multi-scale with similar simple spatial structures.
3) The designed score function is used to calculate the different level of semantic similarity and concatenate the original neighboring features to alleviates the information loss. This processing method allows the model to obtain more reliable semantic augmented local representations.
4) Ablation Study considers very comprehensively, conduct experiments and analysis

from information level, setting of SLA and MGP, and effect of SLA and MGP.

1.6. Open Question

First of all, about the score function in SeA, can we optimize it, such as using other complex calculation formulas (we can read other k neatest neighbors papers for inspiration later). Although this introduces more calculations, it can reduce score mapping errors.

What does the pyramid of this MGP refer to? In this paper, it is a simple merger. Maybe we can re-weight according to different scales?

How to judge that global feature is needed instead of irrelevant semantics? How can we augment the required semantics more specifically(design a new Efficient module)?

1.7. How to relate to our own problem?

First of all, we can consider detecting (or even Re-Identification) multi-scale pedestrians without increasing the receptive field. At the same time, we can consider optimizing the model of this paper under the attention mechanism. Also, on the problem of pedestrian re-identification, we can also filter irrelevant semantics (reduce noise) through the introduce score function. In addition, we can use the paired RGB image as a complement, so that local representations will not have semantic information loss.